MEASUREMENT OF DRIVER PERFORMANCE IN TRAINING SIMULATORS

J.E. Korteling & K. van den Bosch
TNO Institute for Perception
Soesterberg, The Netherlands

<SHEET 1>

One important advantage of training simulators is that they can easily be equipped with an additional system that automatically measures task performance. Training with such a system may provide two major advantages above usual training on a driving simulator: ***more objective an specific performance judgements*** by the instructors *more* ***explicit feedback*** to the student. Therefore, these systems may be called "Performance Evaluation and Feedback systems, in short: PEF systems.

However, in spite of the theoretical advantages of such systems, it seems that these advantages often are not realized. Therefore, in my present talk, I will try to present some principles that are important in the develpment of driver performance evaluation and feedback systems in training simulators.

The principles I will present are largely based on the work we did for the Netherlands Army. For the training of tracked-vehicle drivers (Leopard 2 and YPR) the Netherlands Army has purchased two very expensive full-scale driving simulators. These simulators included, among other things, a computer-generated and collimated image, a six degrees-of-freedom moving-base system and an instruction panel and also a what we call a PEF system.

On the first acquaintance with this PEF system it was noticed that the large quantity of detailed output was not easy to comprehend and seemed to lack significance for driver training. Therefore our institute, the TNO Institute for Perception, was asked by the Netherlands Army to evaluate the system and to give recommendations for improvement.

After this short introduction, I will describe the original version of this PEF system and I will outline some of the shortcomings of this system, just as an example of how you should not make such a system. Furthermore, I will present a design of a more appropriate and user-friendly PEF system. Both the critique of the original and the design of a new system will proceed according to a framework of six steps, or principles, that are crucial for a successful development of automated PEF systems for training simulators.

<SHEET 2, PARTIAL PRINTOUT>

The best way to give you an impression of the old PEF system, as developed by the producer for the Netherlands Army,is to show a part of the printed performance evaluation by this sytem. In order to make this sheet readable, I had to change the

Korteling, J. E., & Bosch, K. van den. (1993). Measurement of driver performance in training simulators. In: Proceedings of International Training Equipment Conference and Exhibition, Wembley, London, UK, May 4-6 (pp. 58-66). Willshire, UK: ITEC Ltd.

layout a little bit, but I don't think this is harmful. I just want to give you an idea of the fact that the whole concept is wrong.

Here you see a pattern of scores of a student on twelve aspects of driving behavior (the upper side), as related to a number of criteria on the same aspects (on the bottom). These criteria actually are the scores of an expert (an instructor). The student's performance on each measure is marked by the degree of similarity to the performance of the expert over the same trajectory. So, for each route-driving measure, the expert database determines the performances leading to a maximum mark. Scores are also weighted. The maximum possible mark, ranges from 2 to 12. Measures regarded as important have a higher maximum mark (e.g., speed: 12) than measures regarded as less important (e.g., gears: 4). Metrics for similarity to expert performance are very arbitrary.

Left-above you see the aspects such as: speed, rotations per minute, degree of acceleration, something about steering wheel rotations, degree of brake pedal push and so forth. In the middle you see the raw scores on the different aspects, and right the marks, as related to the criteria.

Important is that this whole bunch of numbers refers only to 8 sec of driving on a straight road, which ofcourse is only a very short part of a complete trajectory. Therefore the output of a complete trajectory usually is very huge, with printouts exceeding a meter. Here is an example of a complete printout for a student, who drove a coulpe of minutes.

<<SHOW PRINTOUT>>

The sum of the student's marks for all measures within a section of the route (which consists of straight parts, curves, and junctions) or an obstacle (approach, ascent, traverse, descent, or drive off) is expressed as a percentage showing how close the student meets the criteria as produced by the expert. The marks sum to a compound mark of 100% when a student's driving is the same (within the defined minimum ranges) as the expert's driving.

The mean of all section compound percentages over a complete PAM route is called the total mark, reflecting the general similarity of the driving performance of the student relative to the expert's driving behavior. This means that, despite their different length and/or character, section scores are not weighted.

The number of specific problems of this PEF system (that can be identified) is so great that it would take too far, and take to much time, to go into each particular problem. It may be sufficient to state that no instructor working with the simulators wanted to use the system. They did not understand how to wortk with it and could not see its usefulness. Therefore, in the following I will present six principles based on which a PEF system, after my opinion, should be designed. Herewith, also the manner in which this old PEF system does not meet each principle will become evident.

<SHEET, PRINCIPLE. 1> **Objective performance evaluation and explicit feedback should refer to only those subtasks that can be trained with sufficient functional validity.**

This principle is based on the idea that increasing the quality of performance evaluations has no value if the skills that are evaluated on the simulator differ from the skills needed in the real, operational system. In that case, using a PEF system only costs extra time. Therefore, the use of such a system should be limited to those part tasks which are simulated with sufficient validity.

In general the functional validity of the vehicle simulators differs for different subtasks. For example, subtasks that mainly consist of procedures and/or require interaction with artificial parts of the task environment generally allow for more valid simulation than subtasks that require interaction with the natural environment.

This means that the development of a PEF system should start with a description of the task, a task analysis, in which the task to be trained on the simulator is analyzed into its elementary components and subtasks. From this list, only valid trainable subtasks should be selected.

You may then make a list of subtasks that can be trained with sufficient validity. Sometimes you will need experimental research to know that, but often expert-knowledge will be sufficient as a first step.

<SHEET, LIST VALID SUBTASKS>

On this sheet, you see three parts of the list with selected subtasks made. Subtasks have been clustered into three main tasks: route, special actions and obstacle driving. The number of selected subtasks, and thus the lenghth of this list strongly depends on the quality of the simulator with reference to the task to be trained.

<SHEET, PRINCIPLE. 2> **Objective performance evaluation and explicit feedback should refer to the most critical and relevant subtasks of the driving task, while including a broad range of skills necessary for driving performance.**
In order to use a PEF system as efficiently as possible, only the most critical and relevant subtasks should be evaluated. So, this means that the system should not include trivial and/or overlapping subtasks. Also, the total of PEF measurements has to cover a broad range of driving skills as much as possible.

The PAM system in its original form included trivial as well as overlapping subtasks, for example, some obstacles that had to bo crossed were very similar, such that may be evaluated according to the same principles and procedures. Moreover, hardly any of the special actions implied in the training (e.g., slalom course, vehicle clearing course) had been chosen for monitoring.

In order to select key subtasks, we needed the help of the instructors working with the simulator. These people are the real experts concerning the task that has to be trained and their knowledge is indispensible when designing a PEF system.

<SHEET PRINCIPLE 3> **Performance evaluation and feedback should focus on the measures that reflect the most critical aspects of subtasks.**

The third principle concerns the importance of what exactly is measured of a subtask. For example, speed control on a tracked vehicle becomes very critical when driving in sharp curves, whereas this subtask is of secondary importance on straight roads. While in mounting an obstacle, smoothness od driving and speed control are important, while lateral position usually is of minor importance. This means that for different subtasks different critical variables are relevant to represent the quality of driving performance. This issue was not addressed in the original PEF system. In this sytem, the same broad range of variables was measured for nearly every manoever.

Again you have to consult task-experts, such as instructors, to make decisions about this.

<SHEET, **VALID, NON-TRIVIAL/OVERLAPPING, CRITICAL VARIABLES**>

This is the complete list of selected subtasks for the YPR, including their critical variables.

<SHEET PRINCIPLE 4. OBJECTIVE PRINCIPLES> **Performance measures and criteria should be defined according to objective principles, based on characteristics of the vehicle, task analysis, and rules for driving behavior.**

Performance criteria of the old PEF system were based on driving performance of an expert, an instructor, who was supposed to produce the optimal values for the different measures. Ofcourse this is not true. Apart from the variability of the expert's performance, the falseness of this assumption is demonstrated by the fact that many parts of the driving task can be performed satisfactorily using different strategies.

Therefore, the next step is to formulate objective and unambiguous measures and criteria in order to specify exactly what is measured and how it is evaluated.

Knowledge of the vehicle and the driving task offers the best opportunities for this. There usually are objective limits within which the value of variables should be kept, given the driving situation.

For example: in normal situations, you may never drive into the shoulder of the road, rotations per minute while turning on the spot should between: 1500-2000; speed in urban roads: < 30 mph, when turning left or right the direction indicator should be used; and when approaching the step up or the sloping block, driving speed should be decreased until one drives at a foot-pace and these objects should be taken as smoothly as possible.

The relevant measures and criteria may easily be implemented in a new PEF system, such that performance can be judged without the intermediary of an instructor. More details about this point can be found in my paper appearing in the conference book.

<SHEET PRINCIPLE 5: COMPREHENSIBILITY> **Measures, scores and criteria should be easy to comprehend and implications for behavioral improvement should be clear.**

With the original PAM system it was often obscure what exactly was measured. For example, the printouts shows mean and maximum percentages on what is called "steering" or "braking". However, what did such a score exactly mean with respect to driving performance? When a driver has a low score on such an aspect, he has ofcourse to know what he did wrong. When this is unclear, one prominent goal of a system for performance evaluation and feedback is not attained, namely: enhancing the quality of behavioral feedback. Consequently the student still has to improve his driving performance by inefficient trial and error learning.

Furthermore, the interpretation of the heightof marks on the different PAM measures difficult. The prints contained marks that only became meaningful relating them to the expert's marks. When a mark on an aspect is already weighted it does not tell you anything until you have found out how it was weighted.

In order to increase the clarity of scores on a printout, two kinds of indications of the quality of a student's performance may be presented. First, raw scores, but only when these are meaningful in itself. For example: the number of cones hit or the number of gear changes. When this is not the case, (for example when you measure the standard deviation of latteral position, or smoothness of driving in terms af compound accelerations), only transformed scores should be presented indicating the driving performance according to a certain scale. For example: percentile scores, which indicate driving performance relative to ones peer students. When raw scores are difficult to interprete, only these transformed scores should be presented.

Of course, when also total scores are measured, the raw scores have to be weighted because the subtasks and the different measures within a subtask are not always of equal significance. By adding the products of the percentile scores and their weights for all measures, the system can compute total, or compound scores. Because weighing of raw scores will affect the interpretation of total scores, the weights have to be presented clearly on the printout.

<SHEET PRINCIPLE 6, ERGONOMIC DATAPRESENTATION> **Performance data should be presented in a simple and self-explanatory format; irrelevant information should not be provided.**

The output data of the original PEF system were poorly organized, such that it took a substantial amount of effort to get an overview of the performance data. One ergonomic problem in the old system was that each PEF route was divided into many small parts, for each of which performance was evaluated according to the complete set of variables. This resulted in many irrelevant behavioral data per evaluation.

The user friendliness and the effectiveness of a PEF system will improve substantially when the output only contains relevant information.

Thus a print should contain for each subtask one score per measured variable (Possibly extended with a raw score, if this is meaningful).
[[[For each evaluation the instructor should only have to type in the date, the name of the involved PEF database, and the name/number of the student. This information should also be presented on the print.]]]

<SHEET, THE 6 PRINCIPLES>
The original system developed for automated performance measurement had a strong engineering approach. Principles of human factors and psychology concerning performance evaluation and feedback were not taken into account. Therefore, this system showed many problems, ranging from minor shortcomings in the clarity of the output presentation to major flaws in the selection and calculation of appropriate performance measures.

What I did was: formulating six principles, which could be applied in a stepwise manner to develop an adequate PEF system. Here you see an outline of the six principles again.

<SHEET, PRINTOUT NEW PEF SYSTEM>

As a final conclusion of this talk this is a printout of the new designed PEF system. This print shows only subtasks that are supposed to be trained with sufficient validity, No trivial or overlapping subtasks. Measures that are unambiguous critical and meaningfull with respect to the selected subtasks, presented in a self-explanatory format. Left you see the selected subtasks and right the measures belonging to these subtasks, each with a raw score, a percentile score, or both.

If properly implemented, this system would provide a pattern of objective grades on relevant aspects of a students driving behavior, which is easy to comprehend, for the student and for the instructor. Moreover, this system would enhance the feedback to the student (knowledge of results), such that he will get a better idea of what he did well and what he did wrong.

[[[The PEF system as described does not contain criteria for examination. Criteria, or cut-off scores, provide immediate information concerning the question whether or not a student's driving performance is sufficient with respect to specific training objectives. Based on these, it can be decided whether or not a student should be admitted to the next training phases. This kind of criterion may only be implemented after empirical investigation.]]]