Instructional strategy and aptitude in a driving simulator

Martijn L. van Emmerik

Department of Training and Instruction
TNO Human Factors
PO Box 23, 3769 ZG Soesterberg, The Netherlands
vanemmerik@tm.tno.nl

Abstract
In the search for innovative instruction and feedback concepts for simulator training, this paper provides a description of the final experiment of a series of three that was done to compare the effectiveness of verbal versus non-verbal instructions in a driving simulator.

In the experiment, provision of non-verbal instructions was inspired on the principle of cue-augmentation in which naturally occurring cues are emphasized to draw trainees' attention to the proper cues. It has been suggested that this way of information presentation is more intuitive than verbal presentation and therefore might increase training efficiency.

Using the TNO low-cost driving simulator (LOCS), subjects without driving experience received pre-programmed instructions from a professional driving instructor while they learned to drive a simulated car.

The baseline instructions were verbal recordings of 12 basic comments related to speed of driving, position on the road, steering behavior in curves, gear shifting, and negotiating an intersection. In the non-verbal condition the content of each of these instructions was provided as a visual, aural, or haptic cue.

On the basis of their initial performance on a short introductory scenario in the simulator during which no instruction or feedback was provided, each subject was assigned to either the high or low aptitude group. Aptitude was expected to mediate the efficiency of instruction. Both objective simulator data and subjective data (questionnaires) were gathered.

The results showed a large and significant difference in performance between the two aptitude groups. The differences between the instruction conditions were not significant however. Nevertheless, the subjective data showed that both instructor and trainee generally liked the non-verbal instructions. In the low-aptitude group they were clearly used more frequently than the verbal instructions. This may indicate that non-verbal instructions are easier to provide when necessary. If not, they are used just as much as verbal instructions. Based on the comments of the trainees it is hypothesized that non-verbal instruction is less intrusive than verbal instruction. The trainee more often experiences the latter as 'negative' whereas non-verbal instructions are seen as a hint leaving open the way to self-recovery. Therefore, they may be seen as especially helpful to low-aptitude trainees.

Introduction
This paper describes the results of the final experiment in a series of three that was set up to find support for the claim that simulators offer didactical advantages compared to the operational environment. Based on ideas from the literature and experimental findings (e.g. Lintern, Thomley-Yates, Nelson & Roscoe 1987; Roscoe, 1991; Schneider, 1985) it was hypothesized that deviations from reality could enhance training efficiency on a simulator as compared to training on the operational system. A concept that has been used in this connection is augmented cueing (O'Shea, Cook & Young, 1999; Young, Stedmon & Cook, 1999).

Augmented cueing works by emphasizing elements in the (virtual) environment to help trainees focus on the relevant characteristics of a task. This is particularly relevant during the initial phases of training. This way trainees are expected to understand the task faster which in turn helps
speeding up the process of learning. Experimental research on augmented cueing has yielded mixed results, likely as a result from the multitude of aspects affecting instructional efficiency (Lintern & Koonce 1992; O’Shea, Cook & Young, 1999). Clearly, this observation illustrates the relevancy to be aware of possible confounding variables.

Experimental research in the field of training simulation is in fact prone to several sources of confounding. In this light some of the design decisions are explained below.

Although the aim of the experiment was to study the didactical advantages of a simulator compared to the real car it would have been a mistake to test the verbal instruction in a real car and the augmented instructions in a simulator. The differences between the car and the simulator would have introduced a source of confounding making the results inexplicable. As Korteling & Sluimer (1999) point out, it is plausible that a part of the results would have to be attributed to differences in the hardware (car vs. simulator). Which part that would be cannot be determined. This problem was avoided in the present research by using the simulator in both conditions.

Another risk of confounding arises from the differences between instructional content in the conditions. We were interested in the differences between two forms of instruction. If, additionally, differences in content were introduced, this would obscure the experimental results. A fair comparison, therefore, can only occur if the two conditions differ in form only. To assure this, content was kept the same as much as possible. In both conditions instructions and feedback were limited to the same set of actions / mistakes. These were pre-designed, recorded, and assigned to a computer keyboard.

A third possible source of confounding was constituted by ‘a priori skill differences’. Because the data of a previous experiment did not lead to unambiguous conclusions with regard to the differences between the two types of instruction, this follow-up experiment was set up with specific measures to control this variable.

To rule out the effect of a confounding variable, however, subjects should be matched on that variable. A first requisite for this is the availability of a preliminary test that can give a reliable indication of the subjects’ scores on the matching variable. In the present experiment, the relevant question is ‘how can we match subjects on their aptitude for driving the simulator?’ Many standard cognitive and psychomotor tests have been proposed and used to give an indication of driving performance (Ball, Owsley, Sloane, Roenker & Bruni 1993; Heikkilä, 2000). A few examples are ‘visual acuity, contrast sensitivity, eye health, visual memory, personality questionnaires, (choice) reaction time, and information processing tests. Correlation with driving performance (investigating crashing behavior, Ball et al., or looking at faults and offences, Heikkilä) is generally low although studies with specific groups of drivers have been able to yield higher correlations with driving performance. Still it is difficult to generalize from findings referring to specific groups (people with neurological deficiencies or patients suffering from Parkinson’s disease) in order to predict the performance of healthy people who never have driven before. The usability of these measures therefore is questionable. Furthermore, it is known that there are large differences between inexperienced and experienced drivers with regard to their ability to select the appropriate cues and to interpret them correctly (Summala, Lamble & Laakso, 1998).

Method

In this particular experiment subjects were trained to drive a simulated car. They received either verbal instructions or instructions based on cue-augmentation. Both types were activated by keystrokes: in the verbal instruction condition recorded voice-commands were played, and in the augmented instruction condition these commands were translated into a non verbal-instruction (description in Table 1).

An additional aim of this study was to find out if (and how) aptitude differences mediated the effect of the instructions. Therefore, subjects were assigned to different aptitude groups (high- or low aptitude) by means of their performance on a preliminary simulator-driving test. The four groups resulting from this manipulation will be referred to as AI-low or AI-high and VI-low or VI-high (where AI stands for augmented instruction and VI for verbal instruction).

Table 1: Instructions provided in the experiment

<table>
<thead>
<tr>
<th>Key</th>
<th>Type of Instruction</th>
<th>Augmented</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>‘Come on, accelerate’</td>
<td>Roaring engine sound</td>
</tr>
<tr>
<td>2</td>
<td>‘Not faster than 30’</td>
<td>Traffic sign displayed on screen</td>
</tr>
<tr>
<td>3</td>
<td>‘Not faster than 50’</td>
<td>Traffic sign displayed on screen</td>
</tr>
<tr>
<td>4</td>
<td>‘Stop’</td>
<td>Traffic sign displayed on screen</td>
</tr>
<tr>
<td>5</td>
<td>‘Little bit left’</td>
<td>Position line displayed on the road</td>
</tr>
<tr>
<td>6</td>
<td>‘Little bit right’</td>
<td>Position line displayed on the road</td>
</tr>
<tr>
<td>7</td>
<td>‘Attention! Give right of way’</td>
<td>Traffic sign displayed on screen</td>
</tr>
<tr>
<td>8</td>
<td>‘Look carefully at the intersection’</td>
<td>Three beeps (left-right-left)</td>
</tr>
<tr>
<td>9</td>
<td>‘Start turning left now’</td>
<td>Jerk at steering wheel</td>
</tr>
<tr>
<td>10</td>
<td>‘Start turning right now’</td>
<td>Jerk at steering wheel</td>
</tr>
<tr>
<td>11</td>
<td>‘Gear up’</td>
<td>Constant high ‘revs’ sound</td>
</tr>
<tr>
<td>12</td>
<td>‘Gear down’</td>
<td>Constant low ‘revs’ sound</td>
</tr>
</tbody>
</table>

1 Translated from Dutch. All messages were kept as short - and to the point- as possible (6 syllables at most in Dutch).
Instrumentation

The present experiment was conducted in the TNO Low Cost Driving Simulator (LOCS). The LOCS is a PC-based simulator for research into simulator training, validity, transfer, and training effectiveness. It also serves as a demonstrator (see Figure 1 & Figure 2).

The LOCS mock-up provides the interface with the simulation. It consists of a car seat, steering wheel, pedals (brake, acceleration, clutch), gear lever, and an (analogue) speed indicator. All components are original car parts (albeit from different cars) except for the speedometer, which is self made.

Figure 1: The TNO-HF Low-cost simulator (LOCS): five displays are placed in a semi-circle, mirrors are represented by the LCD-screens. The LOCS is equipped with a car seat, steering wheel, pedals for acceleration, brake, and clutch, and a 4-position gearbox.

Figure 2: A closer look at the LOCS. Central monitor above LCD-screen with dashboard information (providing speedometer, odometer, rev meter, and indicator lights). Part of the rearview mirror (LCD-screen) is visible.

Furthermore, the LOCS consists of a number of computers that all serve specific purposes. Three Windows NT machines generate the visual environment on five wide screen (24”) displays that are positioned in a semi-circular configuration. Each of these computer monitors has a horizontal field of view (FOV) of approximately 40° resulting in a total FOV of about 200°. The central screen displays a high-quality image, which is generated by a single computer. The other two computers each generate a reduced resolution image on two peripheral displays. For that reason, the video-card of the central image generator is more powerful than that of the two other machines.

Finally, one MS-DOS computer ('model computer') receives the signals generated in the mock-up (by standard A/D, DAC, and RS232 interfaces) and uses them to compute the corresponding vehicle behavior mathematical model. This machine provides input to the 'sound computer' (Windows '98), which generates the appropriate engine sounds, and communicates with the 'supervisor computer'. The supervisor is used for scenario control, generation of other traffic, and data storage.

The communication between these different computers is established through an Ethernet connection.

Subjects

Twenty-eight subjects participated in this experiment (11 male, 17 female subjects). Seven subjects were assigned to each experimental group. Because the subjects had to be assigned to the groups based on a preliminary test, it was difficult to match subjects on sex as well as on their test results. In the first place, care was taken that the number of high and low skill subjects was equal for both instruction conditions. With regard to the ratio of male / female subjects in each group it can be said that this was correctly represented in both low aptitude groups. Female subjects, however, were slightly over-represented in the high aptitude - augmented instruction group.

The average age of the subjects was 21.1 years with a standard deviation of 2.7 years (youngest 18 oldest 30). None of the subjects had received driving instruction in a car prior to the experiment.

Procedure

In order to match the subjects on their aptitude to drive a simulator it was decided to have the subjects drive two short routes prior to the experiment (approx. 300 meter with 3 curves / intersections). During those test trials no experimental instruction was given. Afterwards the instructor gave an estimation of the ease with which a subject would learn the task. Subsequently the subjects were assigned to one of the conditions (verbal instruction or augmented cues instruction). Subjects did not know what conditions the experiment comprised.

In both conditions subjects received the same four phases of a practice run followed by a test. During practice
the instructor provided help (different depending on instructional condition). During the test no instruction was given (so that the tests were the same for both groups). After the test, the (subjective) instructor judgement was used to decide whether the trainee should continue to the next phase or repeat the same phase.

Each phase introduced some additional difficulties: Initially, the trainees were supposed to drive (at low speed) without shifting gears. No other cars drove around. In phase 2 other traffic was present. Traffic was removed again in phase three but gear shifting was introduced here. Finally in phase 4 trainees had to deal with other traffic while shifting gears.

Figure 3: Experimental design. Diamonds represent a decision of the instructor. VI means 'Verbal Instruction', AI means 'Augmented Instruction'. After the pre-test, all subjects start in phase 1.

Hypotheses
It was hypothesized that differences in skill between groups of trainees would be reflected in the amount of instruction they received during practice, and in their performance during the test trials. Furthermore, we expected subjects that were classified as 'high-aptitude', would perform equally well regardless of the instructional condition they were assigned to. Low aptitude subjects however were expected to profit more from augmented instruction compared to verbal instruction.

Analyses
During the experiment different kinds of data were collected. Prior to the experiment a short questionnaire was administered with questions relating to traffic experience, gaming experience and knowledge of eight basic traffic signs that were to be encountered during the experiment. During practice the amount (and nature) of instructions provided by the instructor was stored for analysis. After the subsequent test the final instructor judgement was recorded (in this phase no instructions were given during driving). During both phases, driving performance was registered with regard to six dependent variables:

- Lateral position, i.e. the distance of the car to the center (white line) of the road.
- Standard deviation of the lateral position (stdevlp), for steadiness of position.
- Speed, in km/h, observing speed limits
- Standard deviation of speed (stdevsp), to register the variability of speed.
- Steering angle, amount of corrections of driving path.
- Standard deviation of steering angle (stdevsta), Smoothness of steering

Practice
Within each practice phase, a comparison of the (relative) amount of instruction between the experimental groups was made with an ANOVA. The differences over the experiment (i.e.) between the phases were not tested statistically because of task differences in each phase. Instead, a qualitative interpretation of these differences is provided.

Test
To evaluate the performance measures, in test phase 1 and 2 a MANOVA was used. An ANOVA was conducted in test phase 3. Because none of the subjects in the category 'augmented / low' participated in the fourth phase, no analyses were done for this test phase.

Results

Questionnaire
Of the 28 subjects, only one was licensed to drive a small motorcycle (<50 cc). None of the subjects had a theory license for car driving although 3 of them planned to take driving lessons within the next 1 to 6 months. Half of the subjects did not have any of such plans at all.

Only 13 subjects played computer games (four of them regularly). Car racing and ‘3D shoot ’em ups’ were both mentioned five times. None of the subjects used a steering wheel while playing.

The number of subjects that was able to name all eight traffic-signs correctly was 11. Twelve persons failed to name two or more signs. (The maximum number of missed or incorrectly named signs was four).
Only in the first phase the ANOVA yielded significant results, $F(3,24) = 3.379$, $p = 0.03$. A post hoc test (Tukey HSD) revealed that the AI low group received significantly more instructions than the AI high group. The differences between the AI-low and the other groups failed to reach significant values, although they were considerably far apart. This may be caused by the fact that Tukey’s HSD is a rather conservative post-hoc test in that it offers a high amount of protection against the increased alpha error rate due to multiple post hoc comparisons. (For example, with the Newman-Keuls or Fisher LSD post-hoc test all three groups differ significantly from AI-low.)

![Graph showing instruction amount per phase for each condition group](image)

Figure 4: **Relative amount of instruction per practice phase for each condition group**

The ANOVAs for the other phases are not significant. Hence, no further tests were conducted.

This same pattern was observed for performance on straight roads and in curves, and also for roads with a speed limit of 50 and 80km/h.

**Discussion and Conclusions**

The present experiment was the final one in a series of three. After the second experiment had been analyzed, it was hypothesized that a confounding variable (aptitude) possibly obscured the (positive) effects of the augmented instruction condition. This (third) experiment was set up to control for differences between subjects as much as possible by matching subjects on their aptitude for learning to drive.

For this means, the instructor was asked to give an estimate of each subject's aptitude based on their performance during a short driving test in the simulator\(^2\). Within that short period of observation, the instructor was able to assign each subject to either the high- or the low-aptitude group. Despite the subjective nature of this procedure, the experimental data (from the following sessions) clearly confirmed the instructor judgment. Yet it proved impossible to distinguish between subject aptitude based on (objective) demographic data, or knowledge of traffic rules as tested in a questionnaire that was also administered prior to the experiment. Contradictory as it may seem, this finding is in line with the results from studies that investigated a number of standard cognitive and psychomotor tests to give an indication of driving performance of neurological patients during the process of rehabilitation (e.g. Ball, Owsley, Sloane, Roenker & Bruni 1993; Heikkilä, 2000). A few examples of such tests are 'visual acuity, contrast sensitivity, eye health, visual memory, personality questionnaires, (choice) reaction time, and information processing tests. None of these has shown a high correlation with 'fitness for driving'. It seems that the best test for 'fitness for driving' is the driving task itself. The use of a driving simulator would allow for standardization of such a test in a safe environment.

With regard to the data it was observed that (just as in the second experiment), the amount of instruction delivered to the augmented group was larger than to the verbal group. On closer examination this is completely caused by the AI low group. This confirms that aptitude certainly is a variable that is capable of obscuring any effects of instruction. As none of the subjects in the augmented group complained about the (high) quantity of the instruction this may indicate that augmented instructions are less intrusive than verbal instructions and therefore may be easier to administer (or receive).

---

\(^2\) Note that this was done prior to the experiment. Subjects had no driving experience and received only a short instruction on how to drive (the simulated car).
Augmented instruction may also be more adaptive than verbal instruction: If there is no need to provide additional instruction, it is used just as much as verbal instruction. This latter type of instruction seems to lack this flexibility. For low aptitude trainees then, it can be a solution to provide non-verbal (augmented) instructions because they seem to be friendlier. The exact (optimal) form of these augmented instruction is yet to be determined.

In previous experiments it was already seen that an error is not the sole criterion for a human instructor to give instruction or feedback. Only on about one out of six possible occasions, an error resulted in an instructor comment (Van Emmerik, De Jong & Van Rooij, 2000). One reason for this selective provision of feedback would be that the instructor would constantly be talking and the trainee would not have time to process the remarks / or direct attention to the (driving) task that is difficult anyway. Another reason may be that the instructor expects the trainee to learn recover from his own mistakes. Verbal instructions can be experienced as a criticism whereas augmented instructions can be seen as a hint leaving open the way to self recovery by the trainee. Therefore, they may be seen as especially helpful to low-aptitude trainees.

A final observation with regard to the augmented instructions was that the amount of AI for the low-aptitude group was smaller than for the other groups only in the third phase of the experiment. This contradicts the assumed flexibility of AI. After all, low aptitude trainees were expected to receive more instruction than their high aptitude counterparts. As a tentative post-hoc explanation it may be suggested that the instructor gave up on these low aptitude subjects. This seems to be supported by the fact that none of the trainees in this group reached the fourth phase.

No significant differences were found in performance in the test phases (even though there was a difference in the amount of instruction). The only difference that was found was in aptitude. Subjects categorized as low aptitude performed worse than high aptitude subjects. This observation, although of no importance to the experimental conditions, confirms the subjectivity of the instructor in judging performance during the experiments.

The interaction between aptitude and condition was never significant during the test phases suggesting that it does not matter what type of instruction (augmented or verbal) subjects received.

Although no definite conclusions can be drawn from these data, a trend was visible showing that low aptitude subjects performed slightly better with verbal instructions. This was something not expected as the augmented instructions were designed to be easily processed and therefore be beneficial in particular to those subjects that were having trouble to perform the task in the first place. Apart from that the (subjective) reactions of trainees were very positive with regard to the augmented cues.

Possibly, this surprising trend is a consequence of the differences between the twelve instructions. While some augmented instructions may have worked very good, others may have been difficult to interpret. This may have given the verbal instructions the overall advantage. Although these speculations could not be tested with this data, they may be an interesting topic for future research.

One thing this experiment may have cleared up though, is the question that remained open after previous experiments: It was suggested that the disappointing results of augmented instruction in that experiment could have been due to the coincidental assigning of low aptitude subjects to the augmented instruction group. This suggestion seems to be confirmed in the final experiment: After the subjects were matched on aptitude prior to assigning them to one of the experimental conditions, no significant differences were found between the instruction conditions (within the groups based on aptitude). The differences between these groups were highly significant however.

It could be true in general that the relatively small effect sizes of experimental manipulations with regard to instructional strategies are easily obscured by inter-individual differences. This is a problem that cannot be solved easily. Experiments of this type could never be done in a 'between subject' design because of the transfer between sessions. One solution probably is to use more subjects and to select them very strictly. Another possibility is to focus on a more detailed level of the driving task. For example, a researcher could pick out one particular instruction to compare different forms of. Both approaches are very laborious. An additional disadvantage of the latter approach is that it would require a task that is more abstract than the current driving task. This would also reduce the validity of the experimental environment and restrict the range of the conclusions.

References


