Post-license Education for Novice Drivers: Evaluation of a Training Programme implemented in Spain.

J. Gabriel Molina
INTRAS, Universitat de Valencia
E-mail: Gabriel.Molina@uv.es

Jaime Sanmartín
INTRAS, Universitat de Valencia.
E-mail: Jaime.Sanmartin@uv.es

Esko Keskinen
Department of Psychology, University of Turku.
E-mail: esko.keskinen@utu.fi

Nick Sanders
CIECA
E-mail: nick.sanders@cieca.be

ABSTRACT

Introduction: This study evaluated the implementation of a 2nd phase training programme for novice drivers in Spain, which puts the primary focus of the training on the higher hierarchical levels of driver behaviour. *Method:* Two hundred and sixty-three participants took part throughout the study, which was implemented as an experimental design with the test and control groups assessed before and after the one day safety training. Measurement of the impact of the training program focused on the participants' self-evaluation and self-reporting of some driving behaviour indicators related to accident risk. *Results:* Data analysis showed a change in the expected direction in the scale related to the skills for careful driving, but not for the other four scales considered. A feedback survey about the training course offered some important input for evaluating the organization, contents, tuition, and results of the three parts of the training programme (discussion group, on-road and track training) as reported by the participants in the test group. *Conclusions and suggestions:* The results of the experiment show that using a one day driver safety course, it is possible to change some of the drivers' evaluations connected to safe driving style into safe direction. The follow-up period was exceptionally long (9 months) and the design (randomly divided experimental and control groups with before and after measurements) was reliable. More effort should be devoted to improving the on-road part of the training, which was often perceived as a typical driving lesson rather than a feedback drive.

KEYWORDS

Post-license Driver Training; Traffic Safety Program; Program Evaluation; Traffic Training; Driver Behaviour.

1. INTRODUCTION

As increased attention is given to road safety issues throughout Europe, policymakers are looking for new ways to address road fatalities, in particular amongst novice drivers. Of the many options available, a 2-phased driver training programme, including a pre-license and post-license phase, is under consideration. Four EU Member States (Finland, Luxembourg, Austria and Estonia) already have obligatory post-license 2nd phase training for all novice drivers. Although the 2nd phase modules differ somewhat from one country to another, the principle is the same: to provide an extended form of driver training which stretches into the initial, high-risk period of independent driving in the months following the driving license.

The EU ADVANCED project (Sanders, 2003) on post-license driver and rider training was commissioned in the year 2000 by the European Union to analyze the current state of voluntary post-license driver and rider training. The logic of this project was that experts and policymakers in countries looking to introduce obligatory 2nd phase training for novice drivers will undoubtedly look to existing infrastructure and resources to facilitate the implementation of such training. However, there are considerable differences in the approach taken in voluntary post-license training and the approach recommended for a 2nd phase for novice drivers. There are also important differences between the approach taken in pre-license training and that advocated for the post-license phase. ADVANCED therefore set guidelines for 2nd phase training and, equally as importantly, also highlighted what to avoid.

The NovEV project (Sanders & Keskinen, 2004) aimed to implement 2nd phase pilot projects across the EU, according to the guidelines laid down in the ADVANCED project. The objectives of the NovEV project were to give the participating countries experience in designing, preparing and implementing such training, and to evaluate the effects of the training on the participants. The aim of this paper is to describe the evaluation results of the implementation of the NovEV project in Spain by Real Automobil Club de Cataluña (RACC), an Automobile Club with close to nine hundred thousand members all around Spain.

1.1. The Training Programme

The 2nd phase training programme implemented in Spain by RACC was based on ADVANCED project guidelines and other scientific studies which put the primary focus of the training on the higher levels of driver behaviour. Research in this area has led to a number of conclusions, including:

- Timing: the training should be spread out over time during the initial period of independent driving, but should allow the novice drivers to have accumulated some driving experience beforehand. (Hatakka et al., 2003).
- Methods: the training should be participant-centred and based on discussion, self-evaluation and self-analysis (requiring considerable coaching and group dynamic skills on the part of the trainers). (Bartl et al., 2002; Hatakka, Keskinen, Gregersen, Glad, & Hernetkoski, 2002).
- Content: the training should above all focus on high level driver behaviour, such as goals for life and skills for living, and the context and motives behind individual car journeys. A theoretical model known as the GDE matrix (Goals for Driver Education) (Hatakka et al., 2002) has been instrumental in developing curricula for driver training at both pre- and post-license stages. The matrix stresses the overriding significance of the higher levels of driver behaviour with regard to accidents and the need for drivers to possess not only knowledge and skills, but also risk awareness and self-evaluation skills at all levels (Hatakka, Keskinen, Katila, & Laapotti, 1997).

RACC's training programme adopted the classic combination of 3 training sessions: track experiences, on-road feedback drives, and group discussion. The track training session consisted of hands-on experience of two simulated driving scenarios: emergency braking with and without ABS on slippery and rough surfaces, for insight training rather than skills training (Gregersen, 1996); and driving with distracting and pressure factors induced. The on-road training session combined urban and rural roads on a pre-defined circuit in which each participant had to drive for 20 minutes as well as to monitor the driving of the other two trainees in the car. After the track and the on-road rides, classroom feedback was programmed for each car team (one trainer and three trainees) in order to discuss about their strengths and weaknesses when driving. In the group

discussion session, the most important sociological and psychological aspects that affect young drivers were introduced and discussed (peer pressure, alcohol and drugs, speeding, risk taking, showing off...).

For largely financial and organizational reasons, the three planned training sessions took place on a single day for each group of trainees. According to the sampling design, seven training days were scheduled: three in Barcelona, three in Valencia, and one in Madrid. A group of 36 people per training day was decided as the optimal number to achieve the three training sessions in the most efficient way. In a regular training day, as people arrived at the training site, they were first received in a welcome introduction, during which the participants were informed about the organization of the training. Then, they were split up into 3 groups of 12 so that each group would be participating in one of the three programmed trainings sessions (on-road, track, and group discussion) at any given time. The three activities took place on a rotational basis, so that all training sessions took place simultaneously but in a different order. In this way, all the three groups had taken two of the three trainings sessions before lunch whereas the remaining one was programmed to be done after lunch.

The aim of the study was to evaluate the impact of this safety training program on young drivers' self-evaluations concerning their risks, skills and habits in traffic. To accomplish this goal, an evaluation strategy was developed, based on an experimental research design with experimental and control groups assessed at two points, before and after the training: firstly, the pre-test should allow to check that the experimental and control groups are equivalent at the outset of the research; secondly, the differences in the post-test for both groups would provide the empirical support to evaluate the impact of the post-license training on the novice drivers.

2. METHOD

2.1. Participants

Subjects included drivers associated to RACC Automobile Club. Three conditions were set to be able to participate in this study: age between 18 and 24 years; no more than 3 years driving experience; and residence in Madrid, Barcelona, Valencia, or towns surrounding these 3 large cities. The last condition was set for logistical reasons associated with the organization of the training. Almost 5800 subjects satisfied the above 3 requirements, which served as the initial database from which were randomly selected 1300 candidates, who were sent a letter offering the opportunity to participate in this research. Finally, 621 subjects agreed to participate: 321 men and 300 women.

The participants were split into control (no training) and test groups according to a random assignment blocked by gender. Some of the 621 subjects who initially joined dropped out during the different steps planned in our research design. In the end, we gained complete records (selection + pre-test + post-test questionnaires) for only 263 subjects, namely 42.3% of the subjects who initially agreed to participate.

2.2. Measures

The measures collected for the evaluation appear differentiated below in three blocks:

(a) Criteria measures. The purpose of the training programme was to encourage safer driving behaviour on the part of the trainees, so indicators that are often connected to drivers who have accidents were considered as the general criteria variables to assess in the pre-test and post-test. Actual accident rates were not considered as criteria measure as long as the study duration and the sample size are not large enough to get reliable indicators of accident occurrence. Thus, we used a set of self-evaluation and self-reported scales related to driving risks, skills and habits. These scales were devised by Hatakka (1998), who describes the development and application of 15 scales designed to measure (un)safe driver behaviour. We opted for 5 of these scales because of their fine psychometric properties for this kind of evaluation. They were included in the pre-test and post-test questionnaires and given to all the subjects in the control and test groups. Thus, the

following five scales were used after being translated into Spanish (note that the short name headings will be used in the rest of the section when referring to them):

- Scale A1: Self-evaluation of risks connected with personal careless driving habits (7 items).
- Scale A2: Self-evaluation of risks connected with showing off and situational reactions (7 items).
- Scale B: Self-evaluation of skills for careful driving (7 items).
- Scale C1: Self-reported habits related to driving while being in an improper state (7 items).
- Scale C2: Self-reported habits related to driving at high speed (8 items).

(b) Control measures. A set of variables usually considered as relevant predictors in traffic research were also collected. These variables enabled us to control the internal validity of the experimental design. They are listed below grouped by affinity:

- About the subject: age; educational level; parents' educational level; years actually driving; possession of own car.
- About the car: car power; estimated car value; car age.
- About driving experience: Km-driven per year; number of fines; number of accidents; type of roads most often used (motorway, national roads, urban); main reasons for driving (going out, to work, on errands, etc.).
- About training interests/motivations: Three scales included in the pre-test indicated the subjects' motivation and reasons for participating in a post-license driving course, according to 3 main areas: increasing driving knowledge and skills; better able to recognize and avoid risks; and improvement of abilities to recognize one's strengths and weaknesses.

(c) Course evaluation measures. Finally, participants in the test group completed a questionnaire, which was included in the post-test, designed to get their opinion on some aspects of the training they took: course organization, trainers' performance, interest in the contents, didactic resources, usefulness, potential improvements, etc. The same questions were raised separately for the 3 modules of the training course: the on-road training, the track training, and the discussion group.

2.3. Procedure

An introduction letter was sent in January 2003 to a near 1300-people random sample of RACC members satisfying the conditions stated in the Participants' Section. The letter described who we were, our aims and work plan, and included an invitation to participate in the study. Recipients were also informed what would be expected of them if they finally decided to participate, as well as what they would obtain by getting involved: besides the intrinsic motivation of improving ones' traffic safety, a raffle with a brand new car as the main prize would be held for all participants in the study, to take place after the post-test collection (March 2004).

The introductory letter also announced that a follow-up telephone call would be made by RACC. This call was made in order to check the letter had been received, to remind recipients of its contents, address any questions and to confirm the participation or non-participation of the subject. If the answer was positive, a set of questions were asked to subjects during their telephone call, in order to collect some initial data on socio-demographic and driving experience variables.

The second mailing (March 2003), to the 621 participants left at this stage consisted of the pre-test, which was completed and returned by 350 subjects. According to the number and residence of the subjects in the test group, 7 one-day training sessions were organized, as follows: 3 in Valencia (July 4th, 5th and 6th), 3 in Barcelona (July 11th, 12th and 13th), and 1 in Madrid (July, 19th). A total of 137 subjects from the test group attended these training sessions that were planned by RACC according to the NovEV project (Sanders & Keskinen, 2004) guidelines. The control and test group participants that satisfied the previous steps received the third mailing with the post-test by December 2003 (9 months after the pre-test), which 263 subjects completed and returned.

After cleaning up the data and running some preliminary analyses, 25 of the subjects' data records were removed from the database because of missing or inconsistent response patterns. The final sample to

consider in the evaluation was thus 238: 114 subjects from the control group and 124 from the test group. Table 1 shows a cross-tabulation by gender and age of the final sample of participants.

Table 1: Cross-tabulation by the research condition, gender and age variables of the final sample of participants.

Research Condition	Gender	Age (reached along 2003)						
		19	20	21	22	23	24	
Control Group	Men	7	9	8	8	7	7	46
	Women	7	10	10	14	11	16	68
Test Group	Men	10	11	9	11	14	5	60
	Women	11	9	7	12	13	12	64
Total		35	39	34	45	45	40	238

2.4. Preliminary Data Analysis

Some preliminary data analyses were achieved in order to check three basic aspects related to the validity of procedure: the characteristics of the sample of participants in relation to the reference population; the quality of the data collected; and the metric properties of the measurement instruments.

2.4.1. Sampling- and self-selection- derived comparisons

Initial data analysis focused on checking how the sampling and self-selection processes had affected the characteristics of the sample of participants involved in the study. For this purpose, we compared the sample of subjects who agreed initially to participate in the experiment (621 of the 1300 who were invited to participate) versus our reference population as described above in the Sample section (N=5728). This comparative analysis (initial sample of participants v. experiment reference population) was possible because some variables were available from RACC database for most of the subjects on the initial list of candidates. Thus, the variables compared for these 2 groups were: age; number of years as driving license holder; number of accidents, car age; and car power. The gender variable was not considered in this comparison as it was blocked when randomly selecting the subjects who were mailed and phoned with the invitation to participate. T-tests for independent samples were computed for each one of these numeric variables. Results showed statistically significant differences only for the number of years having the driving license variable: the sample has had their license shorter than the population (t = 2.86, df = 5591, p = .005). It should be noted, on the one hand, that the difference of the means in both groups (0.11) is equivalent to approximately one month and ten days only, and, on the other hand, that this statistically significant difference should be considered taking into account the large size of the samples compared.

2.4.2. Data quality control

This analysis is directed to detect errors and outliers that can appear in the dataset obtained after data collection and processing. This initial dataset contained 263 data records, each one corresponding to one of the participants who completed all the phases planned in the research design: selection questionnaire, pre-test, training session (only for the test group), and post-test. Four main strategies were applied to clean up the data collected: (1) anomalous values in frequency distributions, cross-tabulations, summary statistics, and uni- and bi- variate graphical representations; (2) subject's very low response rates; (3) extreme standardized means and standard deviations of the subject's response patterns (Dolinger & Dilalla, 1996), and (4) inconsistencies in the subject's expected responses to specific pairs of items. While the first method is variable-oriented and provided information that served either to correct some specific values or to re-code them as missing, the other 3 methods are case-oriented, providing support to locate strange response patterns, whether in the pre-test or in the post-test questionnaires. According to these criteria, data records from 25 subjects were removed from the database so the final sample size in the following analysis was 238.

2.4.3. Psychometric analysis of the measures

This analysis is oriented to check the psychometric properties of the 5 scales used to collect the data that enabled us to test the impact of the training course (see Measures subsection). The five scales were analyzed by obtaining the dimensional structure of the responses given to each scale, in order to test if the original factor structure was reproduced by our data. Dimensional analysis was carried out using the Principal Components Analysis (PCA) model to each one of the response data sets collected for the 5 scales in the pre-test questionnaire (where the number of subjects taking the scales was larger than in the post-test). The results of the factor extraction through PCA showed that the scales A1, A2 and C2 had a mainly unidimensional factor structure, in line with the results of Hatakka (1998) and Keskinen, Hatakka, Katila & Laapotti (1992) with Finnish samples of novice drivers. The Kaiser-Guttman criteria (eigenvalue > 1) was used to set the number of factors to be extracted, which resulted in 2 factors in the case of scales B and C1. However, this second factor in scale B explained a reduced % of variance in relation to the first factor (15.4 v. 35.7), apart from having no sensible interpretation according to the item loadings in the second factor. With regard to scale C1, the mediocre KMO index for this scale (0.67) denotes that PCA is not very well suited to the response data collected for this scale; additionally, the second factor in the scale C1 appears as somewhat relevant if we look at its % of explained variance in relation to that of the first factor (17.8 v. 31.0). However, the low variance accounted for by this second factor (1.25) and the high saturation of all the items in the first factor suggest that it could be assumed that a main factor underlies the response data for this scale, so we cautiously considered it as unidimensional as well. As a consequence of these results, we worked henceforth with the subjects' scores corresponding to the factor scores associated with the first component extracted for each scale.

After analysis of dimensionality, we proceeded with the application of two classical psychometric procedures oriented to evaluate reliability of measures: Cronbach's alpha and the test-retest indexes. The former was obtained for the data collected in the pre-test questionnaire because a larger sample was available than in post-test application and the results were: 0.73 for scale A1; 0.83 for scale A2; 0.68 for scale B; 0.62 for scale C1; and 0.81 for scale C2. With regard to the test-retest reliability index, this was estimated for each scale through the product-moment correlation between the scores obtained from the pre-test and post-test applications of the 5 scales. The calculation of the correlation coefficients was limited to the control group participants as, by not attending the training course, they received less external influences to change than the test group during the period between the pre-test and the post-test. The coefficients obtained were: 0.49 for scale A1; 0.41 for scale A2; 0.40 for scale B; 0.68 for scale C1; and 0.76 for scale C2, clearly higher for the driving habits self-report scales.

3. RESULTS

The statistical analysis proceeded first testing the equivalence between the control and test groups at the outset of the research. Next, the attention was focused on the main goal of this study, namely to evaluate the effectiveness of the training course implemented by RACC according to the NovEV project guidelines (Sanders & Keskinen, 2004). For the latter, a double perspective was adopted according to the two sources of information collected from the participants in this research: driving behaviour indicators related to accident risk (criteria measures); and feedback on the training course they took (course evaluation measures). Note that in the results presented in the following subsections an alpha level of .05 was used for all statistical analyses.

3.1. Equivalence of the test and control groups

A number of analyses were achieved to ensure that the experimental and control groups were equivalent at the outset of the research. With regard to the gender variable, as it was stated above, we blocked this variable in the random assignment of subjects to the research groups. Thus, the distribution of men and women in the initial sample of 621 participants (men: 51.7%; women: 48.3%) is quite similar to this of the final sample in the test group (men: 48.6%; women: 51.6%), however, it is far from this of the control group (men: 40.4%; women: 59.6%), suggesting that the gender variable should be statistically controlled in successive data analysis process.

As was the case for the gender variable, the distribution of the test and control groups could no longer be as balanced as initially expected after the random assignment of the participants to the two research groups. Obviously, the loss of participants during the early phases of the project could have not followed a random pattern. Thus, some checks were carried out to see to what extent the control and the test groups were balanced for some of the variables collected in the selection and pre-test questionnaires. More specifically, the checks covered the criteria (pre-test) and the control variables listed in the Measures subsection. In order to test if these variables were equally distributed in the control and test groups, we proceeded according to the type of variable:

(a) For the category variables, we obtained contingency tables for each one of these crossed by the group variable (control v. test). Pearson's chi-square test was used to contrast the independence of the distribution of the test and control groups with regard to the levels of each one of these variables. The results showed that the control and the test groups were fairly balanced for all the variables considered: all the chi-square values had statistically non-significant results ($p > .05$). Only the education level variable was on the limit of statistical significance ($\chi^2 = 5.98$, df = 2, $p = .050$), a possible explanation for it being that working people (where supposedly more primary and secondary education levels are found) is less favorably disposed to attend a Saturday training programme.

(b) For the numeric variables, t-tests for independent samples were computed for each one of these variables with the research condition as the grouping variable. The null hypothesis of independence resulted in rejection of 6 of the 23 variables. Table 2 summarizes this information for the 6 variables with significant differences.

Table 2: T-tests results for the variables with statistically significant mean differences in the control and test groups in the pre-test phase.

Variable	Research condition	N	Mean	t-test	Cohen d
Car power	Control Group	90	81.01	$t(158) = 2.22$ $p = .028$	0.360
	Test Group	70	73.99		
Mean scores for the 1st scale of training contents interests: Driving knowledge and skills	Control Group	113	3.42	$t(230) = -3.07$ $p = .002$	0.41
	Test Group	119	3.69		
Mean scores for the 2nd scale of training contents interests: Recognizing and avoiding risks	Control Group	113	3.56	$t(229) = -2.07$ $p = .039$	0.27
	Test Group	118	3.76		
Mean scores for the 3rd scale of training contents interests: Recognizing one's strengths and weaknesses	Control Group	113	3.98	$t(230) = -2.99$ $p = .003$	0.42
	Test Group	119	4.21		
Factor scores for the scale A2: Self-evaluation of risks connected with showing-off and situational reactions.	Control Group	114	-0.29	$t(216) = -2.90$ $p = .004$	0.39
	Test Group	119	0.09		
Factor scores for the scale B: Self-evaluation of skills for careful driving.	Control Group	114	0.18	$t(231) = 2.05$ $p = .042$	0.27
	Test Group	119	-0.09		

If we look at the column with *Cohen's d* as estimation of the effect size, we can see that all the values are in the range of what is considered as small values of effect size (< 0.5), denoting that the statistically significant differences found in the analysis could be more related to a sample size factor than to the magnitude of the differences between the 2 research groups. Anyway, given that the effect sizes values for two of the training interest scales are over 0.4, it was advisable to consider these variables in successive analyses. For the sake of simplicity, taking into account that the 3 variables with the scores in the training interests scales work in the same way (having larger values for the test than for the control group), we assumed that a common factor of *Interest in Improving Driving* underlies all three, and we built a summary variable as a mean of them. This new variable would simplify later ANOVA modeling and, as it was expected, the test group (M = 3.89) reported larger interest in improving driving than the control group (M = 3.65) according to the t-test computed for this new summary variable ($t = -3.175$, df = 230, $p = .002$, d = 0.419).

3.2. Main results

Results in this section focus on the specific statistical analysis achieved in order to assess the impact of the 2nd phase training for novice drivers carried out by RACC. This analysis has been achieved on a double basis: (1) the participants' scores in the five scales that were considered as the criteria or dependent variables in this study; (2) the course evaluation, that is, the feedback obtained from participants after the training (see Measures subsection above).

3.2.1. Results related to the criteria measures

An univariate ANOVA model was used to analyze the data of our mixed between-within design for each one of the five scales considered. In all cases, the Group condition (control v. test) was manipulated as a between-subjects factor, while the Time condition was manipulated as a within factor with also 2 levels: the repeated measures obtained for the scale in the pre-test and post-test moments. The hypothesis about the impact of the training course would be supported if, first, the interaction Group by Time (or Group by Time by Gender) was statistically significant, and second, if the analysis of simple effects went in the expected direction: (1) differences between control and test groups in post-test scores; and (2) no differences between control and test groups in pre-test scores.

Prior analysis was carried out to check if the two main assumptions of the ANOVA method that must be upheld in the data were satisfied, namely the gaussian-shaped distribution of the criteria variables, and the homogeneity of the variances of the distributions defined by the levels of the independent variables. Different tests designed to check these two assumptions, like Shapiro-Wilk's and Levene's tests, showed that the normality and homogeneity of variance assumptions were, in general, not satisfied. Exploration of score distributions for the five scales appeared to be quite asymmetrically distributed, which makes sense if we take into account most of the item statements, that is, rather aberrant driving behaviours.

In order to solve this problem, without losing the power of the ANOVA application in the analysis of our data, we opted for the strategy recommended in literature consisting of applying data transformations designed to obtain more symmetric, normal distributions which satisfy the normality assumption and, simultaneously, improve the homogeneity of variances (Berry, 1987; Emerson, 1991). Thus, according to the type and degree of asymmetry in the score distributions for the five scales considered in our study, we applied the most suitable transformation for each case: *SQR* or *LN* for the variables distributed with positive asymmetry; *EXP(2)* for those with negative asymmetry. After applying these transformations on the criteria measures, the acceptance of the hypothesis of normality and homogeneity of variances was corroborated according to Shapiro-Wilk's and Levene's tests, respectively.

Figure 1 contains five graphs that display the mean scores for the results in each one of the scales considered in this study. These shown the scale mean scores for the control and test groups in the pre-test and post-test moments as well as these mean scores differentiated by the gender variable. Just two reminders: (1) the rating scales ranged from 1 to 5 for the five measurement scales applied in this study; (2) positive changes (i.e., positive impact of the training course) were associated to the increase in the subject scores for the 'Skills for Careful Driving' scale, whereas to the decrease of them in the other four scales. The changes that appeared in our study are described below; next, we present the results of testing if these differences were statistically significant.
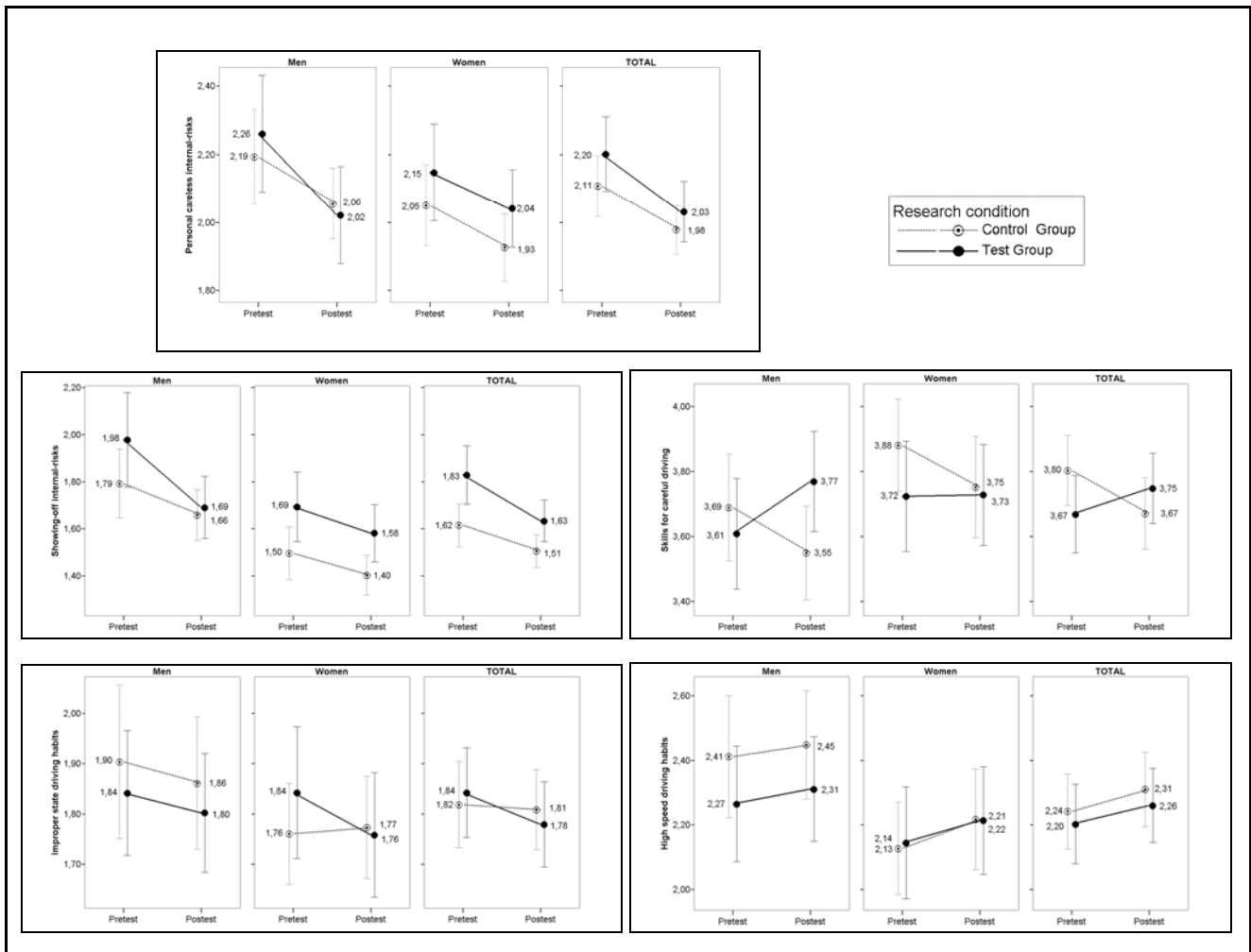
Figure 1: Graphs of mean scores for the five criteria measures considered in this study.

Looking at the total scores in the five scales for the test group in the pre-test and post-test moments, only a negative change appeared in the 'High Speed Driving Habits' scale. It must be noted that this change is rather small (+0.06 scale units) if we take into account the scale reference; moreover, it appeared also in the control group (+0.07 s.u.). Positive changes in the subject scores for the test group were shown in the other four scales:

(a) For the 'Personal Careless Internal Risks' scale is observed a positive change (-0.17 s.u.), however, this is also given in the control group (-0.13 s.u.). A more important effect is shown when considered the men sample (-0.24 s.u.).
(b) For the 'Showing off and Situational Reactions Internal Risks' scale the results are similar to the previous scale. It is shown a positive change in the test group (-0.20 s.u.) and a smaller one in the control group (-0.11), which is more pronounced when considering the young men drivers (-0.29 and -0.13 s.u., respectively). It must be noted the differences between the control and test group in the pretest, which could denote some sample bias derived of the subjects giving away the experiment in any of its phases.
(c) For the 'Skills for Careful Driving' scale, where a score increase is expected, it is found a positive change in the test group (+0.08 s.u.), which is more relevant if we take into account the decrease in the mean scores for the control group (-0.13 s.u.). This difference is much more marked if we look at the men driver sample (+0.16 s.u. and -0.14 s.u., respectively). Again, shocking differences resulted between the control and test group at the pretest moment.
(d) Finally, in the 'Improper State Driving Habits' scale is possible to observe a positive small change for the test group (-0.06 s.u.), yet it is almost null for the control group (-0.01 s.u.). In this case, it is for the women young drivers where this positive change is more important (-0.08 and +0.01 s.u., respectively).

An ANOVA mixed model was applied to the analysis of the data coming from each one of the 5 scales considered in order to check if the previous effects were statistically significant. The gender variable and the scores in the summary variable related to interest in improving driving, which were non-equally distributed in the control and test groups, were also modeled in order to control their potential effect on the dependent variables. Statistically significant effects appeared only for the scale related to the self-evaluation of skills for careful driving (see Table 3). No other higher level interaction effects were found as statistically significant in any case.

Table 3: F-test for the interaction effect (Group by Time) for the five scales considered as criteria measures.

Scale	F	p
Internal risks connected with personal careless driving habits	$F(1,227) = 0.27$	0.60
Internal risks connected with showing off and situational reactions	$F(1,229) = 0.64$	0.42
Skills for careful driving	$F(1,227) = 7.75$	0.01
Driving habits related to being in an improper state	$F(1,227) = 2.27$	0.13
Driving habits related to high speed	$F(1,227) = 0.13$	0.71

According to these results, we proceeded with an analysis of simple effects for the variable with the scores in the 'Skills for Careful Driving' scale. The results of this analysis showed that the source of the significant interaction effect was the difference between the control and test groups in the pre-test moment. In order to control this difference between the two groups at the outset of the research, we reanalyzed the data taking the pre-test scores in this scale as covariate in the ANOVA model. The results of this ANCOVA analysis showed statistically significant differences between the control and test groups in post-test: $F(1.228) = 5.07$; $p = .025$ (see Table 4).

Table 4: ANCOVA results for the 'Skills for Careful Driving' scale.

Source	Sum of squares	Df	Mean square	F	Sig.
Corrected Model	53,66	4	13,42	16,97	,000
Intercept	,14	1	,14	,17	,678
Group	4,01	1	4,01	5,07	,025
Gender	,23	1	,23	,29	,592
FactorScor Pretest	48,02	1	48,02	60,73	,000
Group*Gender	2,19	1	2,20	2,78	,097
Error	180,29	228	,79		
Total	233,98	233			

3.2.2. Results related to the participants' feedback

One hundred and twenty participants from the test group answered the course feedback survey included in the post-test questionnaire. Figure 2 shows the means, and their respective confidence intervals, for the scores in each one of the 10 survey items, which had to be answered for each of the three modules in the course (discussion-group, on-road, and track training). Agreements ratings ranged from 1 ("Not at all") to 5 ("Very much") for all the items. The items appear in Figure 2 ordered from higher to lower mean scores. Finally, note that item #4 ("It was rather boring") was reversed in order this item measures in the same way as the rest of the items. Thus, the descriptive statistics stand for *interesting* rather than *boring* for this item.
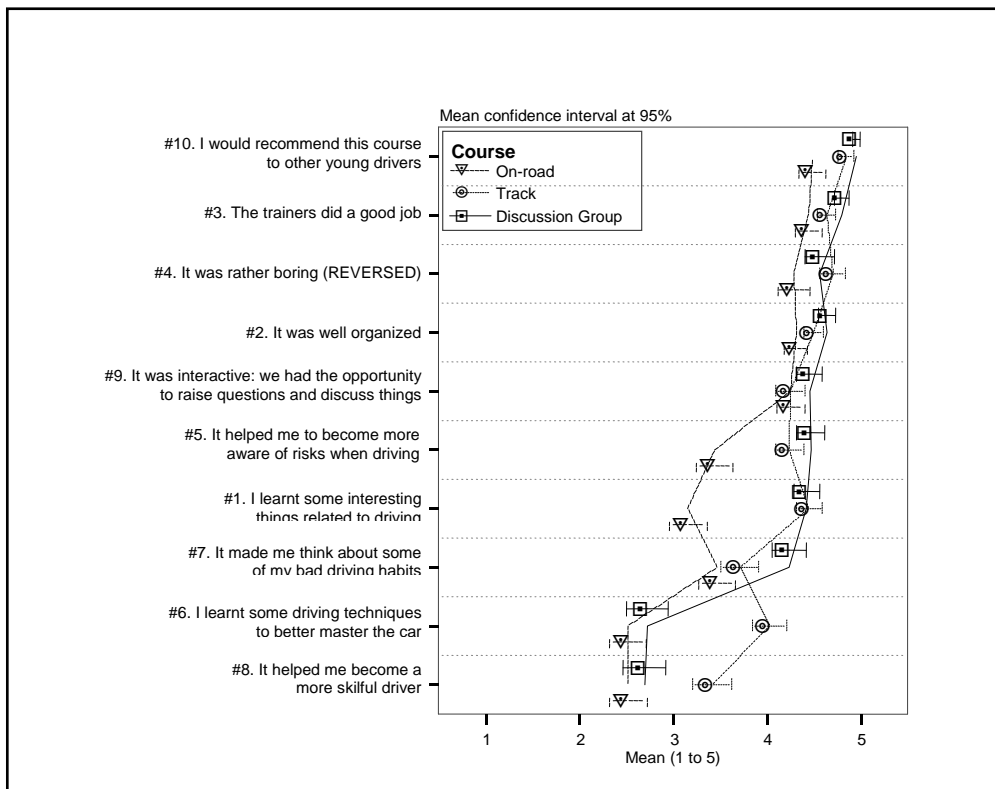
Figure 2: Means and confidence intervals for the items included in the participants' feedback survey.

Item scores for the 10 items in the 3 courses were modeled with an ANOVA mixed model with 2 within-variables (*Item* with 10 levels and *Course* with 3 levels) and the gender variable as between-variable. Given that the latter had no major effect, nor participated with the other variables in significant interaction effects, it was removed from the definite model applied to analyze the data. The ANOVA results showed statistically significant main effects for the Item and the Course factors as well as for the Item by Course interaction, so an analysis of simple effects was computed. The results of this analysis showed that the source of the significant interaction effect is associated to different significant mean differences between the three parts of the course. Many of them are explained by significant lower scores in the items for the on-road part than for the track and group discussion parts (see Figure 2).

4. DISCUSSION

4.1. The impact of training

Main result of the study was that we could find evidence of the impact of safety training after a 9 month follow-up period in a well controlled study. Data analysis results for the five criteria measures considered in this study showed a statistically significant positive change for the 'Skills for Careful Driving' scale. Subject in the experimental group assessed that after 9 month their "Skills for Careful Driving" were still better than before the safety training course. Thus, empirical support was obtained for the impact of the course on the novice drivers' self-evaluation of their skills for careful driving, what goes in the expected direction if we take into account that this self-evaluation facet appears as inversely related to accidents (Hatakka, 1998). It was found a slight trend (not statistically significant) towards expected direction in three of the five scales ("Personal Careless Internal Risks", Showing of and Situational Reactions Internal Risks", and "Improper State Driving Habits"). It is naturally not possible to argue that these criteria measures were improved as a result of the training program. No statistically significant differences were found for the gender variable in any case.

It is interesting to note the observed changes in the control group during the experiment, which probably represents the natural evolution of the young drivers along the 9-month period running from the pre-test to

the post-test times. Another possibility is that these changes have their origin in the measurements made twice as part of the evaluation process. The possible mechanism causing this effect could be that the participants of the study were attending more closely such things in their driving which were asked in the questionnaire. Thus, a positive trend appeared in the control group in two scales ('Personal Careless Internal Risks' and 'Showing off and Situational Reactions Internal Risks'), however, it was not found evidence supporting that the training course caused an improvement added to that natural improvement. Contrary to this trend, a negative one was observed for the 'Skills for Careful Driving' and the 'High Speed Driving Habits' scales in the control group, however, the training course provided a significant change in this trend for the former, yet a null effect for the latter. Finally, a practically null change is observed in the control group for the 'Improper State Driving Habits' self-report scale. Thus, yet the observed trends in the control group were not statistically significant for any scale, the observed trends may suggest which subjects could evolve in a positive way without more intervention that the existing ones, and which could need to be considered as priority intervention targets of training programmes and other traffic policies.

4.2. Participants' feedback

The second line of data analysis in this study, the one based on the participants' feedback survey about the training course, offered some interesting input for evaluating the course as reported by the participants. Mean scores appeared significantly higher for the survey items related to improvements in self-awareness about risks (item #5: "It helped me to become more aware of risks when driving") and bad driving habits (item #7: "It made me think about some of my bad driving habits") than for these related to improvements of driving techniques and skills (items #6 "I learnt some driving techniques to better master the car" and #8 "It helped me become a more skilful driver"), what conforms to the aims of the NovEV project (Sanders & Keskinen, 2004) for such post-license training. It must be noted, however, that survey items #6 and #8 scored significantly higher for the track training, what could indicate an overestimation of the driving skills as a consequence of this training module, something that should be avoided in the potential implementation of the course in future.

With regard to the implementation of the training course, a priority from RACC was to implement a short and feasible training, rather than a long and therefore more comprehensive - but also less realistic from a practical point of view - training. However, the experience in the implementation of the training sessions suggested that each one of the training modules (track, on-road and discussion group) should be extended from 90 minutes to 120 minutes in order to generate better feedback discussions between participants and trainers in each session.

Participants' feedback about the course organization, contents, and tuition revealed a rather positive evaluation of this aspects: items #3 ("The trainers did a good job"), #4 ("It was rather boring [reversed]"), #2 ("It was well organized"), and #9 ("It was interactive: we had the opportunity to raise questions and discuss things.") obtained mean scores over 4 (item rating scales ranging from 1 to 5). It is also outstanding that the general evaluation of the course underlying the statement of item #10 ("I would recommend this course to other young drivers"), which encompasses a comprehensive impression about the training course in the form of an applied consequence obtained the highest item total mean score (4.77) of all the survey items. It should be noted, though, that items #5 ("It [training] helped me become more aware of risks when driving") and #6 ("I learnt some driving techniques to better master the car") showed significant lower scores for the on-road training, so it is advisable to revise the contents and subsequent discussion of this module exercises in order to reduce the possible increase in confidence without increase in skills.

4.3. Training recommendations

Finally, the experience derived of the seven training sessions implemented for this study showed that more efforts should be devoted to improving the on-road part of the training, which was often perceived as a typical driving lesson rather than a feedback drive. According to the experts who were monitoring the NovEV project implementation, it is advisable to consider, at the end of the on-road drive, a better feedback session allowing trainees to raise questions and issues by themselves. It was also recommended that in the group discussion session trainees were encouraged to participate more actively by providing their own experiences. On other hand, some improvements in the research design could have improved the power of

the statistical analysis to detect the potential impact of the training course, for example: (1) a larger sample size; (2) a more balanced sample selection -in our study the control group appeared in the pre-test as systematically *safer* than the test group; and (3) scales with items which were better tailored to the changes expected as a result of the training sessions. As some items were perhaps not very directly related to the course subjects, this may have hampered our attempts to detect specific changes.

ACKNOWLEDGMENT

REFERENCES

Bartl, G., Assailly, J. P., Chatenet, F., Hatakka, M., Keskinen, E., & Willmes-Lenz, G. (2002). [Research Report] *EU-project "Andrea": Analysis of Driver Rehabilitation Programmes.* Vienna: Kuratorium Fur Verkehrssicherheit.

Berry, D. A. (1987). Logarithmic transformations in ANOVA. *Biometrics, 42,* 439-456.

Dolinger, S. J., & Dilalla, D. L. (1996). Cleaning up data and running preliminary analyses. In Frederick T. L. Leong & James T. Austin (Eds.): *The Psychology Research Handbook. A Guide for Graduate Students and Research Assistants.* Thousand Oaks: Sage.

Emerson, J. D. (1991). Introduction to transformations. In *Fundamentals of Exploratory Analysis of Variance.* D. C. Hoalglin, F. Mosteller, & J. W. Tukey (Eds.). NY: Wiley & Sons.

Gregersen, N. P. (1996). Young drivers' overestimation of their own skill – an experiment on the relation between training strategy and skill. *Accident Analysis & Prevention, 28,* 243-250.

Hatakka, M. (1998). Novice Driver's Risk- and Self-Evaluations (Doctoral Thesis). Annales Universitatis Turkuensis, ser.B- TOM.228. Turku, Finland: University of Turku.

Hatakka, M., Keskinen, E., Baughan, C., Goldenbelt, C., Gregersen, N. P., Groot, H., Siegrist, S., Willmes-Lenz, G., & Winkelbauer, M. (2003). [Research Report] *Basic driver training: New models.* Turku, Finland: University of Turku, Department of Psychology.

Hatakka, M., Keskinen, E., Gregersen, N.P., Glad, A., & Hernetkoski, K. (2002). From control of the vehicle to personal self-control; broadening the perspectives to driver education. *Transportation Research Part F. Traffic psychology. 5*, 201-215.

Hatakka, M., Keskinen, E., Katila, A., & Laapotti, S. (1997). Self-reported driving habits are valid predictors of violations and accidents. In T. Rothengatter & E. C. Vaya (Eds.): *Traffic and Transport Psychology, Theory and Application* (pp. 295-303). Amsterdam: Pergamon.

Keskinen, E., Hatakka, M., Katila, A., & Laapotti, S. (1992). [Research Report] *Onnistuiko kuljettajaopetuksen uudistus? Seurantaprojektin loppuraportti.* (Was the renewal of the driver training successful? -english abstract). Psychological Reports n. 94. Turku, Finland: University of Turku.

Sanders, N. (2003). [Research Report] *The EU ADVANCED project: Description and analysis of post-licence driver and rider training* (231 pp.). Amsterdam: CIECA.

Sanders, N., & Keskinen, E. (2004) [Research Report] *The EU NovEV project: Evaluation of post-licence training schemes for novice drivers* (267 pp.). Amsterdam: CIECA.

BIOGRAPHICAL SKETCHES

J. Gabriel Molina, PhD. Dr. Molina is associated professor at the Department of Research Methods in Psychology, University of Valencia, Spain. Most of his research activity has been developed as member of INTRAS (Research Institute on Traffic and Road Safety of the University of Valencia). Main research interests are related to psychological testing, survey methods, and traffic data analysis. He is member of the European Association of Methodology.
Dr. Molina received his doctorate in Psychology from the University of Valencia and he has co-authored some papers published in the journals Applied Psychological Measurement, Educational and Psychological Measurement, Behaviour Research Methods, and Psychological Reports.

Jaime Sanmartín, PhD. Dr. Sanmartín is director of METRAS (Research Group on Measurement, Evaluation, Analysis and Data Processing in Traffic Accidents and Road Safety) at the Research Institute on Traffic and Road Safety of the University of Valencia (INTRAS). He is a member of the foundation team of INTRAS and an active researcher since 1995. He has directed and authored numerous research studies in the development and application of quantitative methods, psychological tests and questionnaires for studying and evaluating human factors in traffic accidents and road safety.
Dr. Sanmartín received his doctorate in Psychology from the University of Valencia, where is associated professor at the Department of Research Methods in Psychology. He is a member of the European Association of Methodology.

Esko Keskinen, PhD. Dr. Keskinen has been working in the field of traffic psychology over 20 years and his special areas are driver education and driver testing as well as accident investigation. He has been an active researcher in several European Union projects concerning driver training and testing and also accident investigation. He has been leading his research group in the area of traffic psychology since 1986.
Prof. Keskinen received his doctorate in Psychology from the University of Turku, Finland. He is professor of Psychology at the Department of Psychology of this University.

Nick Sanders, MA. Nick Sanders is a project manager at the international commission of driver testing authorities (CIECA). He has managed a total of 3 European Union projects in the driver training and testing fields and was recently project coordinator for a 4th. These projects dealt with issues such as novice driver training, driving instructor standards, and medical testing of drivers. Prior to working at CIECA, Mr Sanders was a senior consultant at Deloitte & Touche Europe Services in Brussels, Belgium.
Mr Sanders has a master in European Political Administration from the College of Europe, Bruges (Belgium) and an undergraduate degree in European Studies from the University of Bath, England.