Using non-linear methods to investigate the criterion validity of traffic-psychological test batteries

R. Risser a,*, Ch. Chaloupka a, W. Grundler c, M. Sommer b, J. Häusler b, C. Kaufmann a

a FACTUM Chaloupka & Risser OHG, Verkehrs- und Sozialanalysen, Austria
b Dr. G. Schuhfried GmbH, Mödling, Austria
c Generation Research Program–Humanwissenschaftliches Zentrum der Ludwig, Maximilians, Universität München, Germany

Received 3 November 2006; accepted 19 December 2006

Abstract

In several countries in Europe (among others Germany and Austria) persons who have lost their drivers licence have to undergo a psychological test in order to regain their licence. The article discusses the validity of two test batteries of the Expert System Traffic using standardized driving tests [Schuhfried, G., 2005. Manual Expert System Traffic (XPSV). Schuhfried GmbH, Mödling]. A global evaluation of the respondents’ performance in a standardized driving test was used as criterion measure in order to divide the subjects into drivers, who were classified as relatively safe or unsafe according to their performance in a standardized driving test. Artificial neural networks were used to calculate the criterion validity. This yielded superior classification rates and validity coefficients compared to classical multivariate methods such as a logistic regression. The stability and generalizability of the results was empirically demonstrated using a jack-knife validation, an internal bootstrap validation and an independent validation sample which completed the test batteries and the standardized driving test as part of a so-called traffic-psychological assessment which is compulsory in Austria in all cases, where the driver’s licence has been withdrawn, e.g., when caught driving under the influence of alcohol. Moreover, the procedure allows calculating incremental validities which enable the assessment of the relative importance of the individual predictor variables. This contributes to the transparency of the results obtained with artificial neural networks. In summary it can be said that the results provide empirical evidence of the validity of the traffic-psychological tests batteries used in this study. The practical implications of the results for traffic-psychological assessment are described.

© 2007 Elsevier Ltd. All rights reserved.

Keywords: Fitness to drive; Traffic psychological assessment; Criterion validity; Diagnostical judgment formation; Statistical judgment formation; Artificial neural networks; Expert system traffic

1. Introduction

In many European countries there are legal frameworks regarding traffic-psychological assessment, although their level of elaboration varies (c.f. Kubinger and Litzenberger, 2005). The original intent was to identify drivers at risk to cause accidents or more severe traffic violations. In Austria and Germany traffic-psychological assessment is mandatory to regain a driving licence after a withdrawal due to traffic accidents or severe traffic violations. However, this requires sufficient correspondence between the test batteries used and the respondents’ fitness to drive. Furthermore, the claim that a respondents’ fitness to drive can be predicted assumes, in the first place, that fitness to drive can be measured (Groeger, 1997).

1.1. Measuring fitness to drive

Earlier studies (e.g.: Spoerer, 1965; Undeutsch, 1962) have focused on accident rates as a measure of respondents’ fitness to drive. However, accident rates have several shortcomings. One of these shortcomings has to do with the fact that accidents are rare events and multi-causally determined (Klebelsberg, 1982). This results in unfavorable statistical characteristic such as the Poisson distribution of accident rates. The second shortcoming concerns the measurement of accident rates. The various methods used to measure accident rates, such as data gathered from...
official sources or insurance company records and self-reports, exhibited levels of agreement that varied considerably (11–67; e.g. Owlsley et al., 1991; Szlyk et al., 1995). Furthermore, in comparing self-reports and tape-recorded driving diaries Chapman and Underwood (2000) recently demonstrated that about 80% of the less severe accidents and near accidents were forgotten within a 2 weeks period.

In order to overcome these problems several studies resorted to standardized driving tests as an alternative measure of respondents’ fitness to drive (e.g. Bukasa et al., 1990, 2003; Karner and Neuwirth, 2000; Lundqvist et al., 2000; Sommer et al., 2004). Whenever an alternative measure of respondents’ fitness to drive is proposed, the validity of the alternative criterion is questioned. This would have to be demonstrated by looking at the relationship between the results received with the help of the alternative criterion measurement and the number of accidents subjects’ have been involved in. An analysis of the validity of the Vienna Driving Test was carried out by Chaloupka and Risser (1995). There, the correlation between different accident types for 51 road sections along a standardized route in Vienna (police registered accidents) and behavioral data from the Vienna Driving Test were calculated. The following validity coefficients were derived (Table 1):

The results allow the conclusion that standardized driving tests can be regarded as appropriate criteria for the safety of road users’ behavior (Risser, 1997).

### 1.2. Assessing the criterion validity of standardized traffic-psychological test batteries

Using bivariate statistics, such as correlation coefficients, most recent validation studies of traffic-psychological test batteries yielded validity coefficients which usually did not exceeded .40 (e.g. Bukasa et al., 1990, 2003; Bukasa and Piringer, 2001; Karner and Neuwirth, 2000). However, these methods neither allow the investigation of the incremental validity of the individual tests nor an investigation of the extent to which deficits in one cognitive dimension may be compensated for by other dimensions. More recent studies thus employed multivariate statistical methods to evaluate the criterion validity of traffic-psychological test batteries. Sommer et al. (2004, 2005) and Häusler and Sommer (2006) recently expanded on this idea by suggesting to compare linear and non-linear multivariate statistical methods in order to investigate the appropriateness of the assumptions underlying classical linear multivariate statistical methods.

### 1.3. Classical multivariate methods

Two of the most commonly used multivariate methods to determine the predictive validity of test batteries are discriminant analysis and logistic regression. However, these classical multivariate statistical methods are vulnerable to violations of their statistical preconditions (Bortz, 1999; Brown and Wicker, 2000; Venter and Scott, 2000), and they often lack stability in cross-validation studies (Jahnke, 1982). Furthermore, these classical multivariate methods only model linear or logit correlations between the predictor variables and the criterion and thus assume linear-additive relationships. Interaction or compensatory effects often remain unconsidered.

### 1.4. Non-linear methods: artificial neural networks

One promising alternative to classical multivariate statistical methods is the application of artificial neural networks. An artificial neural network can be conceived as a robust multivariate classification algorithm which assigns respondents to predefined categories based on a set of predictor variables (Bishop, 1995; Dorffner, 1991; Kinnebrock, 1992; Mielke, 2001; Rojas, 2000). The main advantage of artificial neural networks is the ability to model complex non-linear relations (Rakes, 1991). In general, they consist of several elements which are organized into layers according to their function. The input layer represents the predictor variables, while the output layer represents the criterion measure. Between the input layer and the output layer one hidden layer is usually interposed (Kinnebrock, 1992). The elements of the three layers are linked to each other by feed-forward full connections. Thus all elements in a layer transmit information to all elements of the following layer. This structure is commonly referred to as a multi-layer perceptron (see Fig. 1).

The number of hidden layer elements can be freely determined, whereby a larger number of hidden layer elements allow

---

### Table 1

<table>
<thead>
<tr>
<th>Observation variables</th>
<th>Correlation coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Driving extremely on the left or on the right side of the lane</td>
<td>.42 (p = .01)</td>
</tr>
<tr>
<td>Inadequate overtaking</td>
<td>.42 (p = .01)</td>
</tr>
<tr>
<td>Too small lateral distances</td>
<td>.37 (p = .01)</td>
</tr>
<tr>
<td>Delayed lane change in case of obstacles</td>
<td>.38 (p = .01)</td>
</tr>
<tr>
<td>Problems with lane choice (e.g., wrong lane for proceeding after intersection)</td>
<td>.48 (p = .01)</td>
</tr>
<tr>
<td>Exceeding speed limits</td>
<td>.46 (p = .01)</td>
</tr>
<tr>
<td>Distance to the car ahead too short</td>
<td>.29 (p = .05)</td>
</tr>
</tbody>
</table>

---

![Fig. 1. Artificial neural networks incorporate one hidden layer of functional neurons (N1 and N2) between the predictors (P1, P2 and P3) and the criterion (C). The number of functional neurons in the hidden layer determines how many weights need to be optimized and hence the economy of the model.](image)
the modelling of more complex relationships. However, this may also lead to an overfit of the artificial neural network, making it impossible to generalize the predictive model beyond the existing sample (Mielke, 2001). In order to avoid this problem one can determine the number of hidden layer elements by means of the Bayesian Information Criterion (BIC) (c.f. Häusler and Sommer, 2006). Once this has been done the weights of the feed-forward full connections need to be determined using so-called learning algorithms which iteratively optimize the weightings. Once the optimization algorithm reached a certain level of precision, the stability and generalizability of the results obtained with a given artificial neural network needs to be investigated by means of a jack-knife validation, a bootstrap validation and a cross-validation (Häusler and Sommer, 2006). If all three methods indicate that the obtained solution is stable, the researcher may calculate the incremental validity and the relative relevance of the individual predictor variables in order to gain insight into the contribution of the various predictor variables to the overall predictive model (c.f. Häusler and Sommer, 2006). Similar to classical multivariate methods, the incremental validities cannot be interpreted in absolute terms but need to be interpreted in relation to the incremental validities of all the other predictor variables used in the model (c.f. Bortz, 1999; Venter and Scott, 2000).

2. Formulation of the problem

The currently available studies that compare artificial neural networks and classical multivariate methods regarding classification rate and validity coefficients are not entirely consistent. In some studies (c.f. Collins and Clark, 1993, study 2; Somers, 2001; Sommer et al., 2004, 2005; Häusler and Sommer, 2006) artificial neural networks clearly outperformed classical multivariate methods, while these two methods yield similar results in other studies (c.f. Borack, 1995; Collins and Clark, 1993, study 1; Griffin, 1998). According to Bishop (1995), the differences in the results can be explained by a lack of non-linear effects in the data sets of studies, which thus could not indicate a superiority of artificial neural networks. In this case one would expect both types of methods to perform equally well. According to Sommer et al. (2005) the lack of non-linear relations can at least partially be explained by the set of predictor variables used. The authors argued that a more diverse set of predictor variables increases the likelihood of non-linear relations in the data which would lead to a superior performance of artificial neural networks.

3. Method

The data in the here described study were obtained in the frame of a multi-centric approach taken in order to validate two test batteries taken from the Expert System Traffic (Schuhfried, 2005). The data collection took place in Vienna, carried out by Factum OHG, and in Bad Tölz, carried out by the Generation Research Program (GRP). All respondents involved in the study completed the same standardized test battery and took part in a comparable standardized driving test. The data collection was carried out in two separate phases. The data collected in the first phase were used to set up an artificial neural network. In a second phase additional respondents were tested. This second sample is used in order to investigate the generalizability of the results obtained in phase one.

3.1. Samples

The sample in the first phase consisted of 164 (74%) men and 58 (26%) women aged 19–91. The average age was 59 with a standard deviation of 18 years. The median age was 64. A total of 39 (18%) respondents had completed primary school or basic secondary school but without completing vocational training (EU educational level 2), 96 (43%) respondents had completed vocational training (EU educational level 3), 35 (16%) respondents had a qualification at university entrance level (EU educational level 4) and 52 (23%) respondents had a university degree (EU educational level 5).

The sample of the second phase consisted of 40 (80.0%) men and 10 (20.0%) women aged between 18 and 69 years with an average age of 37.3 years and a standard deviation of 12.8 years. The median age was 37.5. A total of 8 (16.0%) respondents had completed compulsory schooling or basic secondary school but without completing vocational training (EU educational level 2), 32 (64.0%) respondents had completed vocational training (EU educational level 3), 10 (20.0%) respondents had a school-leaving qualification at university entrance level (EU educational level 4). All respondents tested in the second phase completed the test batteries and the standardized driving test as part of a so-called traffic-psychological assessment which is compulsory in Austria in all cases, where the driver’s licence has been withdrawn, e.g. when caught driving under the influence of alcohol.

3.2. Measures

3.2.1. Description of the test batteries

The test batteries “Standard” and “Plus” of the Expert System Traffic (Schuhfried, 2005) were used to assess respondents’ cognitive abilities. The two test batteries comprise the following cognitive ability tests:

3.2.1.1. Adaptive matrices test (AMT/S1). This test was administered in order to assess ‘Fluid intelligence’. The items resemble classical matrices, but they are based on explicit construction rules (Hornke et al., 2003). The task of the respondent is to identify the figural pattern among eight answer alternatives which complete the given matrices’ item. The test was administered as a computerized adaptive test. The reliability coefficient of the test amounts to .70.

3.2.1.2. Cognitrone (COG/S11). In order to measure ‘Selective attention’ the Cognitrone test (form S11) was administered. The task of the respondent is to compare an abstract figure with the model in order to assess whether correspondence is given or not. The items are administered without any time limitation. The mean time of correct rejections serves as a measure of ‘Selective
attention’ (Wagner and Karner, 2001). The reliability coefficient amounts to .95.

3.2.1.3. Tachistoscopic traffic perception test (TAVTMB/S1). In this test 20 pictures of traffic scenes are presented to the respondent, for one second each. Then the respondent has to select from a list that contains five different items those ones that he/she remembers to have seen in the picture. The number of correctly answered lists constitutes the main variable ‘Overview’ which serves as a measure of ‘Perceptual Speed’ (Biehl, 1996). The reliability of this variable amounts to .82.

3.2.1.4. Reaction test (RT/S3). This test form measures ‘Decision speed’ and ‘Physical motor speed’ using a simple choice reaction time paradigm (c.f. Schuhfried and Prieler, 1997). Circles of different colors or auditory stimuli are presented to the respondent. During the presentation the finger of the respondent is placed on a rest button. The task of the respondent is to leave the rest button and to press a defined key in case a yellow circle and a certain auditory signal appear in combination. All other types of signals and their combinations have to be ignored. The mean reaction time, defined by the latency from the onset of the relevant stimuli combination to the lifting of the finger from the rest button, serves as a measure of ‘Decision speed’. ‘Physical motor speed’ is measured by the latency from the lifting of the finger from the rest button until the defined key is pressed. ‘Decision speed’ and ‘Physical motor speed’ together sum up to ‘Reaction speed’. The reliability coefficients amount to .94 and .98, respectively.

3.2.1.5. Determination test (DT/S1). This test is used to measure ‘Resilience of attention and reaction speed under conditions of sensory stress’. The task of the respondent is to identify various stimuli and to react to them by pressing the respective corresponding response buttons, using the response panel of the Vienna Test System. The test is administered as a computerized adaptive test whereby the presentation time of the stimuli adjusts itself to the reaction speed of the respondent. However, unlike classic computerized adaptive tests, this test form presents the stimuli a little faster than would be optimal given the respondents’ reaction speed, thus resulting in a condition of sensory stress. In this study the main variable ‘Correct reactions’ was used to assess respondents’ ‘Resilience of attention and reaction speed’ (Schuhfried, 1998). The reliability of the variable ‘Correct’ reactions amounts to .98.

3.2.1.6. Peripheral perception test (PP). This test is used to assess ‘Field of view’ and ‘Divided attention’. Light stimuli move along a panel that is positioned in the periphery of the respondents’ field of vision with a pre-set speed. Whenever such a light stimulus appears the respondent has to react by pressing the foot pedal. At the same time the respondent has to fulfill a central task, which consists in tracking a moving object. Thus, the respondent has to distribute the attentional resources between these two tasks (Schuhfried et al., 2002). The main variables ‘Field of vision’ and ‘Tracking deviation’ are used as predictor variables. Tracking deviation is used as a measure of ‘Divided attention’. The reliability coefficients amount to .96 and .98, respectively.

All the mentioned tests were selected based on previous validation studies. Since administration of the Peripheral Perception Test requires the use of more clumsy equipment that is difficult to carry along, the criterion validity was calculated for the test battery with and without Peripheral Perception. In the following the test battery without the Peripheral Perception Test is referred to as test battery Standard. The test battery that includes the Peripheral Perception Test is referred to as test battery Plus.

3.2.2. Description of the standardized driving test

In addition to completing the above test batteries, each respondent also had to undergo a standardized driving test from which the criterion measurements were derived. The driving test used was the Vienna Driving Test (Chaloupka and Risser, 1995; Burgard, 2005). The standardized driving test was carried out immediately after completion of the tests listed above. Respondents’ performance in the tests was not known to the raters. The observation of driving behavior took place along a route which included a wide range of driving tasks, like lane keeping, lane choice before obstacles and intersections, speeds, distances to the cars ahead and to persons and objects on the road side, overtaking manoeuvres, interaction with vulnerable road users, etc.

The standardized driving tests took about 45 min and were carried out in driving-school cars in the presence of driving instructors. At the end of the test drive the observers made a global assessment of the respondent’s driving behavior. According to Lundqvist et al. (2000), the global assessment of driving behavior is a useful addition to the detailed evaluation commonly recorded in standardized driving tests. This global evaluation consisted of a rating on a standardized five-point rating scale. The same scale is, usually also used for governmental driving tests. In this global rating schema a value of four or five corresponds to driving behaviors which would be insufficient to obtain a driving licence, while 1 is very good, 2 is good and 3 is sufficient. In our study, a driving school teacher and two observers had to assess the driver, and thus average values between the natural Figs. 1–4 or 5 were often received. The three assessing persons were instructed not to communicate about the rating. For the global evaluation of the respondents’ driving behavior we chose a cut-off value of 3.33. In order to reach this figure and even higher values, at least one assessor had to rate with a 4 or a 5. Lower values mean that neither 4 or 5 was given. The inter-rater reliability of the global assessment of driving performance amounted to .86 in Bad Tölz and .95 in Vienna. 60.4% of the respondents in the first phase and 72.7% of the respondents in the second phase were classified as save drivers (<3.33).

4. Results

4.1. Descriptive results

In the first stage of analysis the correlations between the tests of the test batteries Standard and Plus and the global evaluations
of the respondents’ performance in the standardized driving test were calculated. The results can be seen in Table 2.

As can be seen in Table 2, there are significant correlations between the main variables of the two test batteries Plus and Standard and the global evaluation of the respondents’ performance in the standardized driving test. These correlations, however, rarely exceed .40. The comparison of the two samples shows that correlations are for the most part identical, although in relation to the variables ‘Physical motor speed’ and ‘Perceptual speed’ in phase 1 the similarities are weaker.

The variables ‘Field of view’ from the Peripheral Perception Test and ‘Selective attention’ in the Cognitrone constitute an exception. In the sample collected in the first phase, significant correlations were found in both cases, but in the second phase there was no correlation between the global evaluation of the respondents’ driving performance in the standardized driving test and the test variable ‘Field of view’, and there was an inverse correlation between the driving performance and the main variables of the Cognitrone. This can be attributed primarily to changes in the composition of the sample with regard to age. The respondents in the first phase were considerably older than the ones examined in the second phase. One of the main causes of restrictions to vision is aging effects (c.f. Lachenmayr, 2003; Owsley and McGwinn, 1999; Peli and Peli, 2002). The main task in the Cognitrone consists of making decisions about the similarity or non-similarity of complex visual stimuli. Previous studies already demonstrated considerable aging effects in this test variable (e.g. Burgard, 2005; Wagner and Karner, 2001).

4.2. Results obtained with logistic regression

A logistic regression analysis was used to evaluate the predictive validity of both test batteries. The calculations were carried out using SPSS 10.00. In order to ensure the comparability of the results to the ones obtained with artificial neural networks the method “Enter” was chosen. This resulted in a $-2 \log$ likelihood value of 205.30, $\chi^2 (6, N=222) = 23.70, p < .001$ for the test battery Standard and a $-2 \log$ likelihood value of 191.46, $\chi^2 (8, N=222) = 37.60, p < .001$ for the test battery Plus. Table 3 summarizes the validity coefficients and classification rates for the simple classification and for the jack-knife validation for the two test batteries Standard and Plus.

As can be seen in Table 3 the results obtained for both test batteries can be confirmed by the jack-knife method.

The correlation between the classification probabilities of the simple prediction and the jack-knife validation is $R = .97$ for the test battery Standard and $R = .95$ for the test battery Plus, which indicates a good stability of the results. In a further step the two predictive models were validated by means of an internal bootstrap using 1000 bootstrap samples. The validity coefficient and the classification rate were calculated for each bootstrap sample. For the validity coefficient a confidence interval of [.18; .35] and [.22; .37] was obtained for the test battery Standard and the test battery Plus, respectively. This corresponds to a confidence interval for the classification rate of [53.9%; 66.5%] for the test battery Standard and [54.0%; 66.5%] for the test battery Plus.

In a further step the generalizability of the results to the sample of phase two was investigated. This involved using the predictive model of the test batteries Standard and Plus that had been obtained in the sample of phase one for calculating the classification probability of respondents in phase two. The test battery Standard yielded a Spearman correlation of $R = .22$ between the global evaluation of their driving behavior and their classification probability. This corresponds to a classification rate of 54.0%. Both, the validity coefficient and the classification rate are within the confidence intervals of [.18; .35] and [53.9%; 66.5%] obtained in the bootstrap validation. For the test battery Plus the Spearman correlation coefficient amounted to $R = .22$ with a classification rate of 54.0%. Both values are within the confidence intervals of [.22; .37] and [54.0%; 66.5%] obtained in the bootstrap validation. The results thus support the generalizability of the results to new data sets.

Table 2 Spearman correlations between the main variables of the test battery Plus and the overall assessment of driving behavior in phases I and II

<table>
<thead>
<tr>
<th>Sample</th>
<th>Fluid intelligence</th>
<th>Resilience of reaction speed</th>
<th>Decision speed</th>
<th>Physical motor speed</th>
<th>Field of view</th>
<th>Divided attention</th>
<th>Perceptual speed</th>
<th>Selective attention</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phase I</td>
<td>.22*</td>
<td>.34**</td>
<td>-.24**</td>
<td>-.16*</td>
<td>.23**</td>
<td>-.34**</td>
<td>.26**</td>
<td>-.17*</td>
</tr>
<tr>
<td>Phase II</td>
<td>.22*</td>
<td>.33**</td>
<td>-.27**</td>
<td>-.25**</td>
<td>.07</td>
<td>-.33**</td>
<td>.42**</td>
<td>.19</td>
</tr>
</tbody>
</table>

The validity was calculated as the correlation between actual evaluation of the performance in the standardized driving test and the classification probabilities based on the respondents’ test performance.
Table 4
Regression coefficient ($B$), Wald statistic, degrees of freedom (d.f.) and significance level ($p$) for the main variables from the test batteries Standard and Plus.

<table>
<thead>
<tr>
<th>Predictors</th>
<th>B</th>
<th>Wald</th>
<th>d.f.</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AMT: fluid intelligence</td>
<td>.214</td>
<td>1.549</td>
<td>1</td>
<td>.213</td>
</tr>
<tr>
<td>DT: resilience of reaction speed</td>
<td>.011</td>
<td>7.300</td>
<td>1</td>
<td>.007</td>
</tr>
<tr>
<td>RT: decision speed</td>
<td>-.001</td>
<td>7.114</td>
<td>1</td>
<td>.398</td>
</tr>
<tr>
<td>RT: physical motor speed</td>
<td>.001</td>
<td>0.881</td>
<td>1</td>
<td>.776</td>
</tr>
<tr>
<td>TAVT: perceptual speed</td>
<td>.039</td>
<td>0.462</td>
<td>1</td>
<td>.496</td>
</tr>
<tr>
<td>COG: selective attention</td>
<td>.143</td>
<td>0.624</td>
<td>1</td>
<td>.430</td>
</tr>
<tr>
<td>Plus</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AMT: fluid intelligence</td>
<td>.119</td>
<td>0.440</td>
<td>1</td>
<td>.507</td>
</tr>
<tr>
<td>DT: resilience of reaction speed</td>
<td>.009</td>
<td>4.015</td>
<td>1</td>
<td>.045</td>
</tr>
<tr>
<td>RT: decision speed</td>
<td>-.001</td>
<td>0.722</td>
<td>1</td>
<td>.396</td>
</tr>
<tr>
<td>RT: physical motor speed</td>
<td>.001</td>
<td>0.331</td>
<td>1</td>
<td>.565</td>
</tr>
<tr>
<td>PP: field of view</td>
<td>.001</td>
<td>0.008</td>
<td>1</td>
<td>.943</td>
</tr>
<tr>
<td>PP: divided attention</td>
<td>-.132</td>
<td>6.352</td>
<td>1</td>
<td>.012</td>
</tr>
<tr>
<td>TAVT: perceptual speed</td>
<td>.023</td>
<td>0.150</td>
<td>1</td>
<td>.698</td>
</tr>
<tr>
<td>COG: selective attention</td>
<td>.283</td>
<td>2.101</td>
<td>1</td>
<td>.147</td>
</tr>
</tbody>
</table>

After checking the stability and generalizability of the results, the incremental validity of the individual tests of the test batteries Standard and Plus over and above the other variables were calculated. The results are presented in Table 4.

Table 4 indicates that in the test battery Standard only the main variable that reflects ‘Resilience of reaction speed and attention’ (DT/S1) contributes significantly to the overall validity of the predictive model. With regard to the test battery Plus the main variable of DT/S1 as well as the main variable reflecting ‘Divided attention’ (PP) contribute significantly to the overall validity of the predictive model.

4.3. Results obtained with artificial neural networks

The artificial neural network was calculated using the program NN Predict (Häusler, 2004). QuickProp (Fahlman, 1988) was used as the learning algorithm. Following a suggestion of Häusler and Sommer (2006), the number of hidden layer elements was determined by comparing artificial neural networks with varying numbers of hidden layer elements using BIC. This resulted in an optimum number of four hidden layer elements for the test battery Standard and five hidden layer elements for the test battery Plus. Using the empirically derived number of hidden layer elements, the predictive validity of both test batteries with regard to the global evaluation of respondents’ driving performance in the standardized driving test based was investigated. Table 5 summarizes the validity coefficients and classification rates for the simple classification and for the jack-knife validation for the two test batteries Standard and Plus.

As can be seen from Table 5, the results of the prediction of the global evaluation of driving performance in the standardized driving test on the basis of performance in the test batteries Standard and Plus are confirmed by the jack-knife method. The validity coefficients and classification rates, both for the simple prediction and in the jack-knife validation, are high, with a balanced relationship between sensitivity and specificity. The correlation between the classification probabilities of the simple prediction and the jack-knife validation is $R = .98$ for the test battery Standard and $R = .95$ for the test battery Plus. In a further step the two predictive models were validated by means of an internal bootstrap using 1000 bootstrap samples. The validity coefficient and the classification rate were calculated for each bootstrap sample. For the validity coefficients confidence intervals of [.54; .73] and [.61; .79] were obtained for the test battery Standard and the test battery Plus which corresponds to confidence intervals for the classification rate of [74.2%; 85.2%] and [77.5%; 87.2%]. In summary, it can be said that the results of both the bootstrap and the jack-knife validations indicate that the chosen network architecture provides a stable result for both test batteries.

In a further step the generalizability of the results to the sample of phase two was investigated. This involved using the predictive model of the test batteries Standard and Plus obtained in the sample of phase one in order to calculate the classification probability for respondents in phase two. The test battery Standard yielded a Spearman correlation of $R = .55$ between the global evaluation of the respondents driving behavior and their classification probability. The validity coefficient thus lies within the confidence interval of [.54; .73] obtained in the bootstrap validation. A value of 76.0% was obtained for the classification rate; this lies within the confidence interval of [74.2%; 85.2%] obtained in the bootstrap validation. The classification rate is 78.0% and therefore lies within the confidence interval of [77.5%; 87.2%] obtained in the bootstrap validation. The results show that generalizability of the results to new data sets can be assumed.

The next step involved calculating the incremental validity and the relative relevance of the individual test variables of the test batteries.
Table 6 shows that in the test battery Standard ‘Fluid intelligence’ (AMT), ‘Resilience of attention and reaction ability’ (DT), ‘Selective attention’ (COG), ‘Perceptual speed’ (TA VT) and ‘Physical motor speed’ (RT) all contribute considerably to the overall validity of the predictive model. In the test battery Plus, ‘Selective attention’ (COG) and ‘Divided attention’ (PP) have a major contribution to the predictive model (Häusler and Sommer, 2006). Even though the results are rather promising, there are a number of objections often raised to the use of artificial neural networks which should be taken into consideration. The most prominent objection is made on the grounds that artificial neural networks are said to resemble a “black box”, from which the relevance of the individual predictor variables does not follow (c.f. Kimnebrock, 1992; DeTienne et al., 2003). However, the here presented study has shown that this argument does not apply in such a general terms. By comparing models with varying numbers of predictor variables but with otherwise identical network architecture it is possible to calculate at least the incremental validity and the relative relevance which the various test variables contribute to the predictive model (Häusler and Sommer, 2006).

In the present article the criterion validity of two traffic-psychological test batteries was investigated using logistic regression analysis as well as artificial neural networks. The stability and generalizability of the results obtained with these two multivariate statistics proved to be rather satisfying as indicated by the jack-knife validation, the bootstrap validation as well as the independent validation sample. This result is in line with previous studies comparing artificial neural networks and classical multivariate statistics (c.f. Borack, 1995; Collins and Clark, 1993; Somers, 2001; Sommer et al., 2004, 2005; Häusler and Sommer, 2006) and demonstrate that the predictive models can be generalized to a younger sample, which completed the test batteries and the standardized driving test as part of a so-called traffic-psychological assessment which is compulsory in Austria in all cases, where the driver’s licence has been withdrawn, e.g. when caught driving under the influence of alcohol. Furthermore, similar to previous studies using a more varied set of predictor variables artificial neural networks outperform classical multivariate methods such as the logistic regression analysis in the present study. Using artificial neural networks validity coefficients of .68 for the test battery Standard and .78 for the test battery Plus have been obtained while the logistic regression analysis merely yield validity coefficients of .34 for the test battery Standard and .36 for the test battery Plus.

Even though the results are rather promising, there are significant legal consequences for the clients (c.f. Bukasa et al., 2001). Due to the fact that decisions about a respondent’s fitness to drive based on his/her results in traffic-psychological test battery require a sufficient correspondence between the tests results and external measures of driving fitness, the empirical investigation of the criterion validity of traffic-psychological tests is of central relevance. On the basis of theoretical considerations relating to the measurement of traffic safety, global evaluations of the respondents’ driving performance in a standardized driving test were used as criterion measure.
The analysis of the relative relevance of the individual predictor variables revealed that measures of selective and divided attention contribute significantly to the predictive validity of the overall model. This result is in line with prior research on the predictive validity of standardized test batteries in elderly drivers (e.g., Burgard, 2005; De Raedt and Ponjaert-Kristoffersen, 2001; Stutts et al., 1998). Additionally, variables related to perceptual and physical motor reaction speed as well as resilience of reaction speed turned out to contribute significantly to the predictive validity of the test battery in our study. Similar results have been reported by Stutts et al. (1998) using a sample of elderly drivers (>60). The results are, thus, in line with previous research and they add to these results by demonstrating their applicability to a sample of respondents of middle to old age (see sample description). The results also demonstrate the relative relevance of a fluid intelligence measure for the prediction of safe driving behavior. Some initial support for this result can be derived from a study conducted by Daigneault et al. (2002), investigating the predictive validity of several measures of executive functioning such as the Tower of London test in elderly drivers (>65). Given the predictive validity of these measures in the study conducted by Daigneault et al. (2002) and the high correlation observed between measures of executive functioning and fluid intelligence (e.g., Carpenter et al., 1990), one could argue that the results on the relative relevance of the fluid intelligence measure used in the present study are in line with prior empirical studies. However, the results reported in this article also indicate that the relation between safe driving behavior as measured in a standardized driving test and the cognitive dimensions measured by the test battery need not necessarily be a linear one, in contrast to the findings of other studies. This difference is due to the use of an artificial neural network instead of a logistic regression analysis.

However, even though we were able to investigate the relative relevance of our predictor variables in this study, the exact shape of the relationship between the various predictor variables and the criterion measure still remains unknown. Despite these drawbacks, the use of artificial neural networks has several advantages. For instance, the predictive models calculated by means of an artificial neural network allow an empirically validated estimation of the compensation potentials provided by certain ability test variables taken from the traffic-psychological test battery that was used in this study. The person-specific classification probability can be used to estimate the extent to which performance falling short of the required level in one of the cognitive dimensions can be compensated for. However, in practical applications this benefit needs to be restricted to the population of healthy respondents within the age range of the two samples used in the present study. The question whether the predictive model can be generalized to even younger drivers, or, e.g., drivers who suffered a traumatic brain injury, has not been investigated thus far and remains speculative. Furthermore, the authors acknowledge that possible compensatory effects of personality traits (e.g., social responsibility and emotional stability) were not taken into consideration in this study. The practical utility of the results presented in this paper is therefore limited to the estimation of compensatory effects provided by ability test variables (see the interpretation guidelines of the Expert System Traffic; Schuhfried, 2005).

In total, the results indicate that for normal healthy respondents within the given age range the predictive models of the test batteries Standard and Plus with respect to fitness to drive are satisfying, and that the use of artificial neural networks, in addition, allows to provide a thorough assessment of fitness to drive even if there are short comings in certain predictor variables, as the algorithm makes compensation potentials better visible than conventional logistic-regression-based algorithms.

References
