Studia Psychologica, Vol. 59, No. 4, 2017, 256-270 doi: 10.21909/sp.2017.04.746

The External Validity of Psychometric Testing Methods and Their Meaning for Safe Road Performance

Adam Biela, Bohdan Roznowski, Oleg Gorbaniuk, Maria Biela-Warenica
Institute of Psychology, The John Paul II Catholic University of Lublin, Poland

The aim of the paper is to investigate which psychometric tools commonly used by Polish transport psychologists appropriately measure necessary abilities of professional drivers. According to Levis-Evans' differentiation between the driver's performance and the driver's behavior, we explored a statistical relation between the results of tests currently used by transport psychologists, measured according to Szalma's individual differences and safe behaviors on roads. We examine validity of tests using data based on real professional drivers' behavior. The sample included 200 drivers involved in accidents and collisions, and 100 who behaved safely. We tested external validity of chosen psychometric tools by analyzing statistically the relation between test scores and unsafe driving behavior recorded by the police. The results show that only few measurements are valid for differentiation of safe and unsafe drivers. The paper indicates the methodology to reach the prognostic value of the diagnostic tests employed by transport psychologist.

Key words: safety behavior, external validity, transport psychology

Theoretical Background

Since the time when psychologists became interested in road transportation, numerous psychometric tools, including cognitive ability tests and psychomotoric ones have been developed to diagnose individuals applying for a driver's license. This interest in testing these applicants appears already in the classic Münsterberg's book on industrial psychology (Münsterberg, 1913). From this analysis, however, the following question arises: which psychometric tools most appropriately measure abilities that predict safety-related human behavior in road traffic situations?

Szalma (2009) argues that individual differences in human factors/ergonomics should be incorporated into research and practice. Within traffic and transport psychology, different research approaches have attempted to explain individual differences in risky driving behavior and traffic accident involvement, which assumes more complex processes and such factors as: aggressive driving (Galovski, Malta, & Blanchard, 2006; Lajunen, Parker, & Summala, 2004; Ozkan, Lajunen, & Summala, 2006), personality traits (Clare & Robertson, 2005; Nordfjærn & Rundmo, 2013; Ozkan, Lajunen, & Summala, 2006; Oltead & Rundmo, 2006; Sommer et al., 2008), temperament traits (Wontorczyk, 2011), reaction times (Summala, 2000), self-assessment (Sundström, 2008), and the age of drivers (Ball et al., 2005; Clay et al., 2005; Machin & Sankey, 2008; Ulleberg & Rundmo, 2003).

The above listed empirical research studies can be systematized according to the theoretical background developed by Levis-Evans

Correspondence concerning this article should be addressed to Bohdan Roznowski, Institute of Psychology, The John Paul II Catholic University of Lublin, Poland, E-mail: bohdan.roznowski@kul.pl

Received February 24, 2017

(2004), who clearly distinguishes between drivers' performance and drivers' behavior underlying the difference between them. The behavior seems to be determined by the driver's performance and his or her subjective estimation of a traffic situation (e.g., road width, Levis-Evans & Charlton, 2006), based also on the driver's motivation and risk perception (Levis-Evans & Rothengatter, 2009). In our analysis we will specify which performance factors and personality factors will be considered as prognostically most relevant to the driver's safety behavior.

The above literature shows that there is no single perspective on the complex map of factors that determine drivers' safety behavior but, rather, that the factors and tests are taken separately to predict risky driving behavior. This sets the ground for the need to seriously question whether the tools used by transport psychologists have sufficient psychometric status in predicting road safety behavior. Transport psychologists have been using tests that are already in practice or they have constructed new ones, theoretically assuming that they measure the abilities crucial for human safety behavior on public roads (Gorbaniuk et al., 2015). Thus, the following question can be asked: whether such an intuitive and theoretical – constructional validity is really sufficient for predicting drivers' safety behavior?

The most relevant external criterion for the efficiency of professional road drivers seems to be the safety behavior on public roads (McCormick & Tiffin, 1980). In spite of the objections related to employing traffic accidents (Svansson & Hyden, 2006), we claim that the analysis of traffic conflicts requires creating artificial situations for drivers and that is why it is connected with higher measurement error than the real road traffic environment. We can begin our reflection by considering how to define the validity of tests that have been used by transport psychologists. In doing so, we will follow some critical remarks of Sartori and Pasini (2007), who raise the issue of: how we can be sure that psychological tests measure what they are assumed to measure?

Bertua, Anderson and Salgado (2005) show that numerous meta-analytic studies in the USA and Europe indicate high validity of cognitive ability tests across different occupational groups, where drivers were one of the tested groups. These studies operationalize the predictive validity of tests in terms of criterion-related-validity, where this type criterion constitutes job behavior in a standardized situation and training success. Similar research findings were reported by Sommer et al. (2008), who aimed to discover the validity of cognitive abilities and personality traits in predicting different aspects of traffic safety (Oltedal & Rundmo, 2006). The results demonstrate the incremental validity of selected personality measures in predicting standardized driving test performance.

However, all the studies noted above, understand the validity of the tests used by transport psychologists as a relationship between the outcomes of the relevant test, obtained by the drivers, or the candidates for drivers, and their efficiency assessment in a standardized driving test taken under experimental or training conditions (Lincoln et al., 2010). Therefore, the considered validity of transport psychology tests, in effect, deals with standardized driving test experiments or training conditions, which is important but not sufficient, rather than with the criterion related to real-life safety driving on public roads which, in turn, are very different situations. Such a statement is not in accordance with the classical industrial psychology methodology, which requires comparing test outcomes with real-life behavior in a perspective of two years (e.g., McCormick & Tiffin, 1980; Landy & Conte, 2010).

It is noted by some authors (e.g., Risser et al., 2008; Sommer et al., 2008) that the criterion validity of traffic-psychology test batteries requires a sufficient correspondence between the tests' results and external measures of public road driving behavior. The stability and generalizability of the obtained results proved to be rather satisfying as indicated by the jackknife validation, the bootstrap validation, and the independent validation sample. However, there still remains a question: what kind of criterion validity of traffic-psychology test batteries can we accept as a satisfactory one?

While trying to answer this question, we assume that safety road behavior criterion such as police recordings can differentiate the drivers who were involved in traffic events from safe drivers and seems to be one of possible external criterion for assessing external validity of psychometric tests used by transport psychologists. In traffic psychology literature traffic events are frequently divided into two categories: traffic conflicts and accidents (Svensson & Hyden, 2006). Two models of the relation between traffic conflicts and accidents are presented by Güttinger (1982; Laureshyn et al., 2017). For the reason underlined above, measuring traffic conflicts demands creating standardized conditions, which creates an artificial situation for testing drivers, we focus only on accidents.

In terms of the severity level, the accidents are often divided into five categories: property damage only (PDO), possible injury, nonincapacitating injury, incapacitating injury, and fatal injury (Al-Ghamdi, 2002; Kaplan & Prato, 2012) or three levels: property damage only, injury, and fatal (Abellán et al., 2013; De Lapparent, 2006; Zhang et al., 2013). The Highway Safety Manual (2010) provides two levels of severity: fatal-and-injury (FI) or property damage only (PDO). That is why in our paper we suggest using only the two-categories: 1) collisions causing property damage only and 2) accidents including injury and fatal situations. We hypothesize that the drivers, who have caused an accident will obtain worse results than drivers who do not have a traffic police record.

This article aims to present, firstly, the methodology for a real measure of driving-safety-behavior on public roads, which employs traffic police records as the external validity related criterion for the psychometric tools used by transport psychologists, and then, secondly, to indicate which methods proved to be valid, according to this methodology. In our analysis we will also incorporate the issue of external validity as a methodological criterion for evaluating occupational safety intervention research (Shannon, Robson, & Guastello, 1999), and also the multidimensional approach in organizational behavioral research proposed by Edwards (2001).

Method

Measurements of Independent Variables

As the independent variable in our research, we selected nine tests that are typically used by transport psychologists in Poland to diagnose the competencies of professional road drivers (i.e., visual perception skills, personality traits, mental abilities, locomotoric skills). Since there are no state regulations as to which particular psychometric tools should be used in Poland, each traffic psychological test center is free to choose from the range of available tests for diagnosing a profile of competences and abilities required for drivers.

The reason behind choosing all the available tests was to determine their differentiating explanatory power in terms of road safety behavior. For example, extraversion and neuroticism with NEO-FFI measures and Eysenck's Test, and also every series of Raven's tests were selected because they measure similar but different abilities of analogical and inductive reasoning (Costa & McCrae, 1992; Raven, Raven, & Court, 2000).

These tests were systematized into six groups of dispositions-measuring psychometric tools as shown in Table 1. The particular tests vary in the number of scores used as the outcomes for testing each individual. The total number of test scores reached in testing, using all nine tests for each individual subject, was 27. The numbers and symbols of particular test scores are listed in Table 1.

Table 1 *List of measurements used in the study*

No	Name of the measurements
	Visual perception accuracy tests
1	Stereoscopic Vision Test
2	Dark Room Test: vision in the dark
3	Dark Room Test: sensitivity to glare
	Mental ability tests: Raven Progressive Matrices Test
4	Series A
5	Series B
6	Series C
7	Series D
8	Series E
9	Test Sum of all series
	Attention tests: Poppelreuter Tables Test
10	the longest series correctly recorded any numbers
11	number of mistakes made in series of numbers written
12	total number of correctly written numbers
	NEO Five Factor Inventory by P. T. Costa, R. R. McCrae (NEO-FFI)
13	Neuroticism (NEO-FFI NEU)
14	Extraversion (NEO-FFI EXT)
15	Openness to experience (NEO-FFI OPN)
16	Agreeableness (NEO-FFI AGB)
17	Conscientiousness (NEO-FFI CON)
	Eysenck Personality Questionnaire Revised (EPQ-R)
18	Neuroticism (EPQ-R N)
19	Extraversion (EPQ-R E)
20	Psychoticism (EPQ-R P)
21	Social Desirability (EPQ-R L)
	Tests of loco-motoric abilities
22	Reaction Time Meter: simple reaction time
23	Reaction Time Meter: distribution of simple reaction time
24	Reaction Time Meter: complex reaction time
25	Reaction Time Meter: distribution of complex reaction time
26	Reaction Time Meter: mistakes of complex reaction
27	The Piórkowski Apparatus for measuring complex eye-hand coordination
28	Kinestezjometr Apparatus for measuring kinesthetic sensitivity and the precise movement of the legs

Unsafe Road Behavior Scale (URBS) – the dependent variable

While elaborating on definitions of severity dimensions, the authors (Laureshyn et al., 2017) distinguish between the risk intensity of injury (accidents) and the risk intensity of collision. This fits the behavioral approach of our research, where we differentiate between the two rank levels of unsafe behavior of drivers. However, for the method of our study we suggest that it is important to treat drivers, who took part in traffic events without injuring others but where material damage occurred (what we call *collision*), differently from those who took part in traffic events with more serious consequences, for example, where people were injured or even killed (in our paper this is referred to as an *accident*). Moreover, we think that differentiation should also be made between drivers who caused accidents or collisions and those who did not cause but were involved in a road accident or a collision.

Thus, the dependent/explained variable in our study was the driver's safe road behavior in opposition to common practice with self-reported driving pattern assessment (Ozkan, Lajunen, & Summala, 2006). This means that we have secured external behavioral criteria, i.e., safety level of road drivers' behavior. Moreover, in the analyzed cases the external criteria have an evident inter-subjectively controlled verification due to the routine road police analysis and classification of the drivers' behavior records, i.e., as having caused the road accident or collision, or being involved in circumstances related to an accident or a collision.

The drivers' behavior is systematized in accordance with the following five-rank order safety behavior scale: 1. A1(a), 2. A1(b), 3. A1(c), 4. A1(d), 5. A2, i.e., from the riskiest behavior group of drivers to the group with the safest behavior. This rank order scale assumes the conjunction of two behavioral criteria: 1) to cause or not to cause such traffic events, where the driver becomes a perpetrator or a victim of accident or a collision and 2) weight of consequences of the traffic events – killed or injured persons (an accident), or those with only material consequences (a collision). It should also be underlined that participants of a traffic situation might also include drivers who avoid, by an evasive action or by a chance, an accident or a collision in this situation.

The first rank concerning the safety road behavior scale is allocated to the drivers who caused the accident as confirmed by the police records [A1(a)]. These drivers caused a motor vehicle accident in which one or more vehicles were involved as well as other road users participating in a traffic event and some people were injured or killed – what we call an accident. The second rank is assigned to the drivers who, according to the road police records, caused a collision [A1(b)]. They were the perpetrators of a motor vehicle traffic event in which one or more vehicles were involved but nobody was killed or suffered injuries. The third rank is assigned to the drivers who were not the perpetrators but the victims of a road accident [A1(c)]. This means that they were not the perpetrators but the victims of motor vehicle accidents where someone was injured or killed. The fourth rank is assigned to the drivers who were also not perpetrators but victims of traffic event where nobody was killed or suffered injuries [A1(d)]. In such cases the only consequences of the traffic event comprised material damage. Finally, the fifth rank is assigned to the drivers who had actually participated in the traffic event situation but avoided, by their evasive action, an accident or a collision, and as a result did not appear in road police reports (the control group A2). All in all the rank scale is interpreted as the Unsafe Road Behavior Scale (URBS).

However, at this point one can question the reliability and validity of the URBS scale. This is very central to our methodology, as it has to be stated that the URBS measures human behavior of our subjects on public roads as the dependent variable. In fact, the reliability of the dependent variable measurement in our study was limited to the period of the last two years prior to the psychological study. This is because of the Polish law, according to which professional drivers must undergo an obligatory psychological testing every 2 years. However, the professional drivers who participated in road accidents are ordered by the police to pass psychological examinations irrespective of the timing of the periodic examinations. Hence the information on being the perpetrator or the victim of a road accident is fully reliable: it is reported via 100% of the cases in the police registers and it is confirmed by the court in case of any dispute over who is at fault.

Participation in traffic collisions does not require police intervention if there is only minor material damage and the parties come to an agreement in terms of who is responsible for causing the damage. Such cases are not registered in police records. On the other hand, if the material damage is large or the parties cannot reach an agreement, then the case is recorded by the police, and in case of any doubt it is settled by the court. If the police intervention concerns a collision, the driver is legally obliged to undergo a compulsory psychological testing. Thus, all the people in our study who qualified as collision participants were actually participants in a collision confirmed by the police information and have not caused any traffic accident in the last two years. On the other hand, the drivers in our control group have certainly not caused any road accidents in the last 2 years, nevertheless, some individuals in the control group are likely to have been involved in a low-harm/material collision but hid this fact from the psychologist.

Subjects and Statistical Analyses

The tested drivers were differentiated according to the police public road “unsafe behavior” records. The four risky groups of 50 drivers each (200 professionals total), mentioned above, with a history of a safety-related problem (A1a-d), were randomly selected from the subject pool of professional drivers sent for an obligatory testing by their employer and included in each group discriminated in terms of the URBS. The recruitment of 100 subjects for the control group (A2) is based on the periodic rudimentary psychometric diagnosis of professional drivers employed by companies as part of the legal requirement in Poland. Transportation psychologists who were licensed to diagnose road drivers tested the participants individually in a standardized way at the Psychological Center for Drivers in Biłgoraj (Poland).

The subjects were aged between 21 and 65. The data concerning the number of subjects belonging to the particular safety behavior groups and their age descriptive statistics are collected in Table 2.

The descriptive statistics presented in Table 2 shows that the analyzed groups of drivers in our research project are, unfortunately, not homogenous in terms of age of the drivers. We considered age as a meaningful factor for drivers' behavior (Feng, Li, Ci, & Zhang, 2016; Schaie, 1994; Thompson et al., 2012). The carried out analysis clearly showed that the compared groups of subjects are statistically significantly differentiated ($F = 3.52, p < .01, \eta^2 = .049$). Therefore, we decided to use the age of the tested drivers as a covariant factor and employed one-way covariance analysis (1-ANCOVA aimed at answering the following question: is belonging of the tested drivers to a particular group of road safety behavior a statistically significant source of variance, as far as the analyzed outcomes of the diagnostic measurements are concerned? ANCOVA ex-

Table 2 *Number of participants and their age*

The groups of drivers to their URB criterion	Number of drivers	Main descriptive statistics concerning age of the subjects	
		M	*SD*
A1(a)	50	35.84	11.66
A1(b)	50	37.72	10.68
A1(c)	50	41.44	10.21
A1(d)	50	39.50	9.35
A2	76	42.21	10.73
Total	276	39.61	10.76

amines the influence of the independent variables on the dependent variable (safety traffic behavior) while removing the effect of the covariate factor (age). ANCOVA first conducts a regression of the covariate on the dependent variable. The residuals (the unexplained variance in the regression model) are then subject to ANOVA.

The most important factor for this study is the difference between groups A2 and A1(a). We used the planned contrast t-test and Cohen's d to analyze whether the drivers from two opposite groups are differentiated by the particular measurement. The result of this analysis is also supported by the accounted correlation ratio between every group located on the rank scale and the measurements (τ-c).

Results

Table 3 shows the main descriptive statistics reached by the tested drivers of the control group and the particularly risky behavior groups (according to the traffic police records) on the psychometric tests used (as listed in Table 1).

Which psychometric tests can predict risky behavior of drivers based on the five-point rank order scale?

Moving on to the next stage of our analysis, we establish the relevance between the outcomes of these tests and the position of the drivers on the five-point rank order behavioral scale, i.e. their road traffic safety record. This is why our analysis should be extended to the differences between the groups of drivers who scored above risky on the behavior scale. The one-way ANCOVA outcomes show that 11 out of 27 considered test scores reached a statistically significant level in differentiating five groups of drivers under consideration (see Table 3). They are as follows: mistakes of complex reaction ($F(4,295) = 4.92$; $p < .001$; $\eta^2 = 0.068$), distribution of complex reaction time ($F(4,295) = 4.21$; $p < .001$; $\eta^2 = 0.059$), and complex reaction time ($F(4,295) = 4.34$; $p < .01$; $\eta^2 = .060$). All of the above mentioned test outcomes concern the Reaction Time Meter. In terms of sight effectiveness, statistical significance indicators were reached in: sensitivity to glare ($F(4,295) = 6.23$; $p < .001$; $\eta^2 = .085$) and vision in the dark ($F(4,295) = 4.24$; $p < .01$; $\eta^2 = .059$). The other significant factors were found in the following measures: complex eye-hand coordination ($F(4,295) = 4.75$; $p < .05$; $\eta^2 = .066$), the longest series of correctly remembered numbers ($F(4,295) = 2.68$; $p < .05$; $\eta^2 = .038$), and resolution of figures into constituent parts ($F(4,295) = 2.83$; $p < .05$; $\eta^2 = .040$).

However, the outcomes presented above require more in-depth comparative analysis of the

Table 3 *Descriptive statistics of all variables and compared groups of subjects*

Diagnostic tests and indicators	Safe road behavior ($n_1 = 100$)		Involved in collision ($n_2 = 50$)		Involved in accident ($n_3 = 50$)		Caused collision ($n_4 = 50$)		Caused accident ($n_5 = 50$)		One-way ANCOVA		Planned contrast test		Kendall's Tau-c Rank Correlation
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*F*	η^2	t^b	*Cohen's d*	τ-c^b
Stereoscopic Vision Test	.400	.102	.402	.178	.408	.112	.522	.587	.577	.693	2.16	.031	1.43	.25	.02
Dark Room Test:															
• vision in the dark	11.25	2.15	11.42	2.23	12.55	3.74	12.61	5.36	14.04	5.82	4.24**	.059	3.04***	.53	.09*
• sensitivity to glare	9.92	2.04	9.76	1.86	11.01	4.07	10.96	4.48	12.99	5.52	6.23***	.085	3.89***	.68	.13**
Raven's Progressive Matrices, Test Series:															
• A: Noticing continuous patterns	11.45	0.74	11.48	.81	11.70	0.62	11.44	.61	11.34	.75	1.77	.026	-.31	.05	.00
• B: Noticing analogies between pairs of figures	10.76	1.34	10.98	.98	10.59	1.54	10.51	1.42	10.31	1.97	1.53	.022	-1.50	.26	-.06
• C: Noticing progressive alterations of figures	9.05	1.31	9.06	1.33	9.05	1.46	8.65	1.63	8.96	1.66	.72	.011	.28	.05	.02
• D: Noticing permutations of figures	8.91	1.56	8.82	2.08	9.08	1.71	8.45	1.86	8.59	2.11	.98	.014	-.13	.02	.01
• E: Resolution of figures into constituent parts	3.76	2.53	4.69	2.43	4.10	2.28	3.22	2.06	3.72	2.51	2.83*	.040	.98	.17	.02
• Sum: Logical induction	43.93	4.71	45.00	5.60	44.47	5.25	42.38	5.61	42.87	6.12	2.19	.031	-.11	.02	.01
Poppelreuter Tables Test:															
• the longest series correctly written numbers	24.96	4.08	23.18	6.49	23.40	5.32	21.75	6.10	22.41	7.16	2.68*	.038	-1.93*	.34	-.11**
• total number of correctly written numbers	25.45	4.16	24.78	5.22	25.02	4.24	23.16	5.47	24.11	5.52	1.84	.027	-.97	.17	-.06
NEO-FFI:															
• Neuroticism	12.72	5.40	14.17	5.12	14.15	5.80	14.03	4.72	15.37	5.75	1.93	.028	1.40	.24	.05
• Extraversion	31.24	4.68	30.31	5.41	30.18	5.04	29.42	5.20	28.80	6.40	1.82	.026	-1.72	.30	-.07
• Openness to experience	23.01	4.04	22.76	4.22	22.83	3.86	23.05	4.89	22.58	4.48	.11	.002	-.42	.07	-.03
• Agreeableness	34.68	3.85	33.66	4.09	33.31	5.39	32.43	4.18	32.38	4.17	2.91*	.041	-2.20**	.38	-.12**
• Conscientiousness	37.37	5.37	37.63	4.49	36.47	4.47	35.65	3.99	35.10	5.68	2.61*	.037	-1.86*	.32	-.10*
EPQ-R:															
• Neuroticism	3.51	2.60	4.26	3.06	4.50	2.89	4.54	3.27	4.86	3.51	1.77	.026	2.41***	.42	.12**
• Extraversion	15.39	3.46	13.87	4.05	14.80	3.74	15.22	3.84	14.98	3.93	1.39	.020	-.20	.04	.01
• Psychoticism	4.27	2.72	4.96	2.08	5.11	3.15	4.87	2.04	5.33	2.53	1.53	.022	2.45***	.43	.12**
• Social desirability	16.55	3.63	14.79	4.85	14.73	3.69	14.67	4.33	13.92	4.67	3.46**	.049	-3.14***	.55	-.16***
Reaction Time Metter:															
• simple reaction time	.244	.021	.245	.017	.246	.021	.250	.025	.252	.024	1.29	.019	1.93*	.34	.06
• distribution of simple reaction time	.112	.039	.107	.031	.118	.043	.113	.036	.128	.053	1.80	.026	1.88*	.33	.07
• complex reaction time	.394	.028	.401	.033	.402	.026	.413	.036	.417	.047	4.34**	.060	2.86***	.50	.16***
• distribution of complex reaction time	.278	.044	.279	.056	.283	.051	.295	.075	.320	.082	4.21**	.059	4.14***	.72	.15***
• mistakes of complex reaction	1.07	.76	1.26	.88	1.17	.77	1.65	1.22	1.75	1.31	4.92**	.068	3.62***	.63	.18***
The Piórkowski Apparatus[a]	90.20	5.58	88.53	9.32	88.98	7.07	84.92	10.86	84.20	12.81	4.75**	.066	-1.99*	.35	-.09*
Kinestezjometr Apparatus	2.08	0.48	2.06	.45	2.27	.48	2.11	.44	2.11	.51	1.59	.023	.27	.05	.23***

Note. a – test *F* and *t* performed for logarithmized variable; b – one-side test
$*p < 0.05$; $**p < 0.01$; $***p < 0.001$

mean scores of particular tests for each grade of the rank order scale of risky road behavior of the drivers surveyed. Figure 1 illustrates the standardized means of the psychometric test, which differentiated statistical significance in ANCOVA outcomes. The drivers belonging to the drivers' groups differed in terms of safety road behavior on the rank order 5-point scale. This illustration shows that only the outcomes of the Raven's Progressive Matrices Test, Set E (resolution of figures into constituent parts) manifests a curved-line regularity, which does not transfer into a validity measure for this test in order to estimate the exposure to dangerous behaviors of drivers as recorded by the traffic police. In case of the mean results of the other tests, we have to deal with dependencies that can be approximated to a straight-line relationship. That is why the outcomes of the above tests can be used to forecast drivers' risky vs. safe behaviors as inspected and recorded by road police.

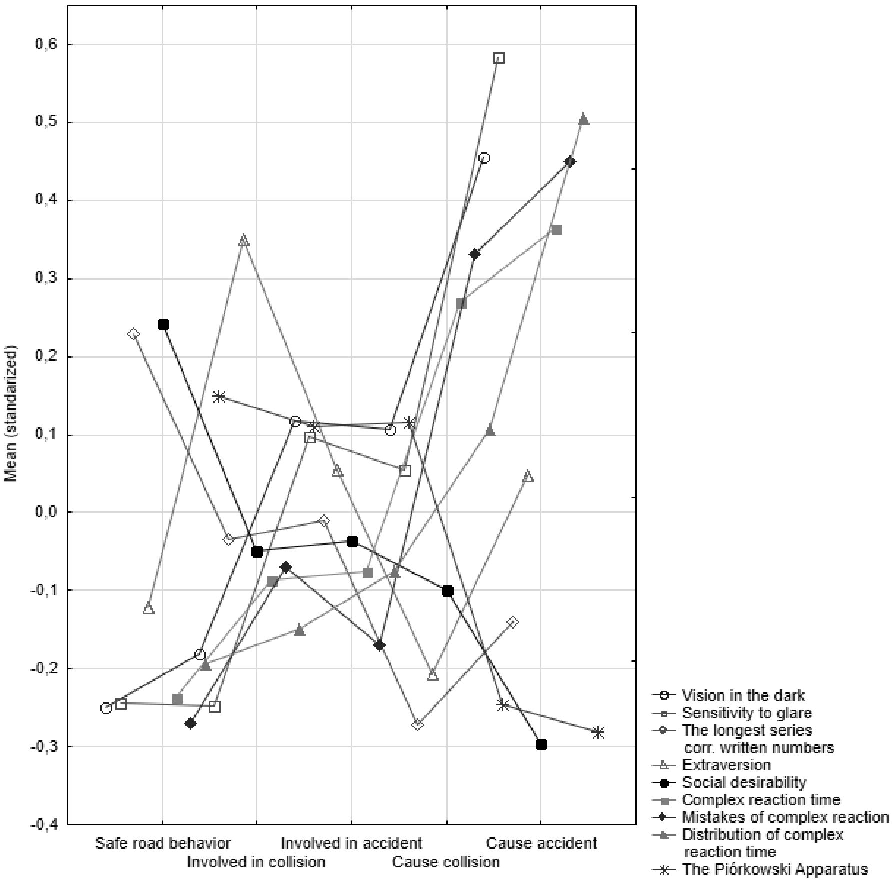

Figure 1 Means of test in particular groups of five-point rank scale of safety vs. risky traffic behavior

To what extent can some tests predict drivers' behavior of the most fatal consequences?

The most serious threat on public roads is drivers who cause traffic accidents that result in injury or death. Therefore, we decided to look for tests that are most reliable in predicting this relationship. We have checked how far the means in the group of drivers who caused an accident (i.e., the highest risk group) differ from the means of drivers who have no road police record (i.e., the control group) (see the appropriate columns in Table 3).

The test results differentiating the most risky drivers from the drivers who have no road police record are as follows: Dark Room Test (DRT): vision in the dark ($t = 3.04$; $df = 298$, $p < .001$, $d = .53$), DRT: sensitivity to glare ($t = 3.86$; $p < .001$, $d = .68$), Poppelreuter Tables Test: the longest series of correctly written numbers ($t = -1.93$; $p < .05$, $d = .34$), Agreeableness (NEO-FFI) ($t = -2.20$; $p < .05$, $d = .38$), Conscientiousness (NEO-FFI) ($t = -1.86$; $p < .05$, $d = .32$), Neuroticism (EPQ-R) ($t = 2.41$; $p < .001$, $d = .42$), Psychoticism (EPQ-R) ($t = 2.45$; $p < .001$, $d = .43$), Social desirability (EPQ-R L) ($t = -3.14$; $p < .001$, $d = .55$), Reaction Time Meter: simple reaction time ($t = 1.93$; $p < .05$, $d = .34$), Reaction Time Meter: distribution of simple reaction time ($t = 1.88$; $p < 0.05$, $d = .33$), Reaction Time Meter: complex reaction time ($t = 2.87$; $p < .001$, $d = .50$), Reaction Time Meter (RTM): distribution of complex reaction time ($t = 4.14$; $p < .001$, $d = .72$), RTM: mistakes of complex reaction ($t = 3.62$; $p < .001$, $d = .63$), The Piórkowski Apparatus for measuring complex eye-hand coordination ($t = -1.99$; $p < .05$, $d = .35$). In all of the above listed tests, the standardized differences between means are higher than $d > .30$, which demonstrates the powerful nature of these tests.

As both criteria, i.e. position on a five-point rank scale and the result of comparison of extreme groups, should be taken into account. We considered only the tests the outcomes of which were previously shown to be significant in the one-way analysis of variance. In this case we established eight measurements: mistakes of complex reaction, distribution of complex reaction time, complex reaction time, sensitivity to glare, vision in the dark, complex eye-hand coordination, the longest series of correctly recorded random numbers, and desirability.

Correlational approach to external validity of psychometric tests

The final part of our analysis will be a correlational approach, which is a classic methodology employed to assess the external validity of psychometric tests used in our research. The dependent variable in our research is the five-point rank order risky vs. safety behavior scale. The independent variables (psychometric tests) are expressed on the appropriate cardinal scale – we have decided to employ the Kendall's tau-c coefficient for our analysis since we consider it more suitable for rectangular tables. In this analysis the test outcome of each individual driver will be referred to the location of the subject on the five-point rank order safety behavior scale. We have noted 11 statistically significant (among 27 accounted for) but relatively low correlations. For our validity analysis we only accepted the correlations with p-value less than .01.

The highest correlation τ-c = .18 was reached by the score of mistakes of complex reaction as measured by the Measure of Reaction Time. This coefficient points to a higher indication of mistakes of complex reactions of the tested drivers, the more serious the consequences in risky road behavior of the drivers. A similar trend and strength of relationship is also shown in the results measured by the Reaction Time Measure that deals with complex reaction as such. The length of complex reaction time has a positive correlation of τ-c = .16 with the rank order scale of the road risky vs. safety behavior of drivers. Similarly, the spread of the complex reaction time responses is associated with more dangerous drivers (τ-c = .15). The same is true as far as the sensitivity to glare (measured by Dark Room Test) is concerned. Lower sensitivity to glare is linked with the lower safety behavior of drivers.

The correlation between the level of psychoticism and neuroticism of drivers (measured by EPQ-R) and risky road behavior is also statistically significant. Drivers with a higher level of psychoticism and neuroticism are more at risk of being involved in traffic events (τ-c = .12).

Negative correlations are found between the risky behavior scale of the drivers and the following psychometric tests scores: Social Desirability measured by EPQ-R (τ-c = .16), the scale of Agreeableness as measured by NEO-FFI (τ-c = .12), and the longest series of correctly recorded numbers as measured by the Poppelreuter Table Test (τ-c = .11).

Discussion

This part of our paper attempts to answer the following questions: 1) What is the post-factum external validity of the psychometric tests used by transportation psychologists in some countries to diagnose abilities for safe driving? 2) Which psychometric tools should be recommended in predicting drivers' safe behavior on public roads? 3) How far are the particular measures of the psychometric tools, used by transportation psychologists when testing drivers, really able to predict the risky behavior of drivers as recorded by road police assessments? And finally, 4) How can we advance the post-factum validity of diagnostic tests into their predictive validity?

What is the post-factum external validity of psychometric tests in road behavior?

In order to be more precise in our forecasting we should point out the object of predicting. In our analysis we considered two kinds of forecasting based on particular test outcomes: 1) The one that most contrasts the two risky behavior groups of drivers, and 2) forecasting the five-rank order risky behavior groups.

Considering forecasting the most contrasting of the two risky behavior groups of drivers, 14 out of 27 tests used by transportation psychologist have been found to be valuable predictors of one's driving behavior, which could be registered (or not) in road police records of accidents with death or injury consequences. Analyzing the forecasting of the five-rank-order drivers' risky behavior informs us that in 11 out of 27 scores, the measure has the particular ability of discriminating risky behavior among drivers. It means that when a driver attains a higher score (or lower score in some measures), it allows us to state that this driver will get (or has already received) a road police record, placing him in the higher rank order position of the risky behavior scale, in contrast to the driver who reached a much lower score in the test.

However, the analysis of results illustrated in Figure 1 allows us to state that only one score (i.e., the Raven's Progressive Matrices Test, Set E) does not manifest dependencies that can be approximated to straight-line relationships. This means that the outcomes of the other eight tests can be used to predict drivers' risky behaviors as inspected and recorded by road police.

Which psychometric tests should be chosen for psychological predicting?

The above presented discussion allows us to conclude that all of the psychological tests used by transportation psychologists have some theoretical validity or implicit rationality based on practical psychological experience. However, our research and, particularly, employing the five-point rank scale of drivers' risky behavior founded on road police records, gives us the possibility to indicate some psychometric tests, which could be recommended for use by transportation psychologists. The proposed conjunctive criterion for such recommendation is a level of statistical significance not lower than of $p < .01$, to underline reliability, concerning: 1) one-way ANCOVA in differentiating the particular measure outcomes of the tested drivers in accordance with their five-point safe behavior scale, and 2) level of τ-c correlation coefficient between the test outcomes of the individual drivers that participated in the research and the rank of the risky behavior scale of these drivers, i.e. our URBS.

According to these criteria the following four tests would be recommended for use in transportation psychology practice (see Table 4): 1) Dark Room Test – vision in the dark, 2) The Poppelreuter Tables Test – the longest series of correctly recorded numbers, 3) Reaction Time Meter – complex reaction time, and 4) Eysenck's Personality Questionnaire Revised (EPQ-R) – measuring Social Desirability (EPQ-R, L).

The first two tests belong to psychometric tools that measure cognitive abilities (and, particularly visual perception accuracy abilities) and attention required in safe road behavior. The most predictive factor of drivers' safe behavior appeared to be the scores of tests measuring abilities such as sensitivity to glare in darkness and the score of the longest series of correctly recorded numbers.

The third recommended test-score of high post-factum external validity belongs to the tests measuring the loco-motoric abilities of drivers. Among these tests, the most valid one was found in the scores of RTM, and in the scores of the complex reaction time: higher level of these scores is required for safe behavior while driving. Thus, we finally decided to choose the measure of failure of complex reaction time for predicting safe behavior because the interpretation of its test score seems to be the most predictable indicator of safe behavior of drivers.

The fourth recommended test-score of high post-factum external validity is the Social Desirability scale score (L) of the EPQ-R with high level increasing safe behavior of drivers on public roads.

Is *post-factum* (i.e., ex-post) external validity a stage towards predictive validity?

We have to realize the fact that the above discussed external validity analysis reflects only the methodology used for stating the pre-

Table 4 *Tests recommended for use in transportation psychology practice*

No	Name of diagnostic measurement
	Visual perception accuracy tests
3	Dark Room Test: sensitivity to glare
8	Mental ability tests: Raven Progressive Matrices Test Series E
	Eysenck Personality Questionnaire Revised (EPQ-R)
21	Social Desirablity (EPQ-R L)
	Tests of loco-motoric abilities
24	Reaction Time Meter: complex reaction time
25	Reaction Time Meter: distribution of complex reaction time
26	Reaction Time Meter: mistakes of complex reaction

dictive validity of diagnostic tests for drivers (called as post-factum or ex-post validation). The starting point of this methodology is to interpret the severity of unsafe drivers' behavior in terms of their intensity impact on a public road situation. We are proposing to measure this intensity impact on the rank order scale, which has been operationalized above as URBS. Therefore, the DT-UB correlation can be interpreted as an *ex-post* (or *post-factum*) unsafe road behavior validation of the diagnostic test for drivers. However, it can be considered as a natural methodological and psychometric stage to reach the prognostic value of these tests. In order to determine the prognostic validity of the employed diagnostic tests, the required step should involve assessing the same drivers in terms of their URBS at least two years after their psychometric testing day conducted by a transportation psychologist.

Research Limitation and Future Research Directions

In the following research, only accidents and collisions were taken into consideration, although traffic accidents are very rear events and also very complicated. Since Hyden (1987), the so called phenomenon of "Safety Pyramid" has been known with its base of undisturbed passages, which are very safe and occur most of the time. The very top of the pyramid consists of the most severe events such as fatal or serious injury accidents. Svensson (1990) explores more explicitly the top of Hyden's pyramid in terms of the "severity diamond's" model. Moreover, it should also be underlined that among participants of any road situation there might also be drivers who avoid, by an evasive action or by chance, an accident or a collision in a given situation.

Measuring the dependent variable in our research has some other limitations as well. Firstly, the established correlations of safe behavior on the roads measured by the dependent variable refer to the period preceding 2 years prior to the psychological testing. It is possible that respondents participated in more serious accidents or traffic collisions in the period beforehand. Secondly, some people in the control group in the past 2 years might have participated in less serious road traffic accidents and have not reported this to either the psychologist or the employer.

The number of subjects is not high, but collecting even such a number was difficult bearing in mind a wide range of measurements used in the analysis. It is a common practice to use only few of them in a diagnosis. Future research could take into account extending the number of subjects, especially those tested in many different Psychological Centers for Drivers.

Another limitation pertains to the measurements, which are only used in Poland. Future research would benefit from involving the international battery of tests used in the field of testing drivers, e.g., the Vienna Test System, which would provide an international sample. Generally, transportation psychology authors recommend a particular diagnostic test or test battery to diagnose abilities for safe road performance. Research is conducted on personality traits (Nordfjærn & Rundmo, 2013; Ozkan et al., 2006; Oltead & Rundmo, 2006; Sommer et al., 2008), self-assessment (Sundström, 2008), temperament traits (Wontorczyk, 2011), and cognitive abilities (Bertua et al., 2005).

The uniqueness of our recommendation is that we based our analysis on the whole spectrum of psychometric tools, which are used by transportation psychologists in Poland (Gorbaniuk et al., 2015) and the external criterion is drivers' real behavior.

This research diagnosed the kind of abilities, skills and other competences profiles of tested drivers are of good enough prognostic value for predicting real life drivers' safe behavior on public roads, measured via URBS recorded by road police. Of course, each test employed in our research is based on the defined theoretical background concerning human potentialities for behavior in the defined situation (cognitive abilities: accuracy of visual perception, mental ability, attention skills, personality traits, locomotoric abilities). The recommended tests for drivers belong to the above mentioned three groups of tools measuring the abilities, skills and competencies for safe performance on roads.

Conclusions

In the following research, the relationship between driving performance (psychometric measurements) and driver's behavior (police evidence) was tested. The external behavior in our analysis was the real-life drivers' risky behavior as assessed and documented by road police in their records. The obtained results suggest that the use of transportation psychology psychometric tests in diagnosing professional drivers has a theoretically and empirically justified validity. Our study allows us to recommend four measurements for transportation psychologists, which explain drivers' safe behavior on public roads: 1) sensitivity to glare in darkness, 2) attention, 3) failures in complex reaction time, and 4) social desirability as a personality characteristic. The presented research can be considered as an important stage to reach the prognostic value of these measurements.

References

Abellán, J., López, G., & De OñA, J. (2013). Analysis of traffic accident severity using decision rules via decision trees. *Expert Systems with Applications*, *40*, 6047-6054.

Al-Ghamdi, A. S. (2002). Using logistic regression to estimate the influence of accident factors on accident severity. *Accident Analysis & Prevention*, *34*, 729-741.

Ball, K. K., Clay, O. J., Wadley, V. G., Roth, D. L., Edwards, J. D., & Roenker, D. L. (2005). *Predicting driving performance in older adults with the useful field of view test: A meta-analysis.* Proceedings of the 3rd International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design, Rockport, Maine, June 27-30.

Bertua, C., Anderson N., & Salgado J. F. (2005). The predictive validity of cognitive ability tests: A UK meta-analysis. *Journal of Occupational and Organizational Psychology*, *78*, 387-409.

Clare, S., & Robertson, I. T. (2005). A meta-analytic review of the Big Five personality factors and accident involvement in occupational and non-occupational settings. *Journal of Occupational and Organizational Psychology*, *78*, 355-376.

Clay, O. J., Wadley, V. G., Edwards, J. D., Roth D. L., Roenker, D. L., & Ball, K. K. (2005). Cumulative meta-analysis of the relationship between useful field of view and driving performance in older adults: Current and future implications. *Optometry and Vision Science, 82*, 724-731.

Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) manual.* Odessa, FL: Psychological Assessment Resources.

De Lapparent, M. (2006). Empirical Bayesian analysis of accident severity for motorcyclists in large French urban areas. *Accident Analysis & Prevention*, *38*, 260-268.

Edwards, J. R. (2001). Multidimensional constructs in organizational behavior research: An integrative analytical framework. *Organizational Research Methods*, *4*, 144-192.

Feng, S., Li, Z., Ci, Y., & Zhang,G. (2016). Risk factors affecting fatal bus accident severity: Their impact on different types of bus drivers. *Accident Analysis & Prevention, 86,* 29-39.

Galovski, T. E., Malta, L. S., & Blanchard, E. B. (2006). *Road rage: Assessment and treatment of the angry, aggressive driver*. Washington, DC, US: American Psychological Association.

Gorabniuk, O., Rożnowski, B., Biela, A., & Biela-Warenica, M. (2015). What is measured by the psychometric tools used for driver aptitude assessment in Poland? A research report. *Annals of Psychology*, *18*, 145-155.

Güttinger, V. A. (1982). From accidents to conflicts: Alternative safety measurement. Third International Workshop on Traffic Conflicts Techniques, Leidschendam, The Netherlands, April 1982.

Highway Safety Manual (2010). Washington, D.C.: American Association of State Highway and Transportation Officials.

Kaplan, S., & Prato, C. G. (2012). Risk factors associated with bus accident severity in the United States: A generalized ordered logit model. *Journal of Safety Research*, *43*, 171-180.

Lajunen, T., Parker, D., & Summala, H. (2004). The Manchester Driver Behaviour Questionnaire: A cross-cultural study. *Accident Analysis & Prevention*, *36*, 231-248.

Landy, F. J., & Conte, J. M. (2010). *Work in the 21st Century*. Danvers: J. Wiley.

Laureshyn, A., Jonsson, C., Ceunynck De, T., Svensson, A., Goede de, M., Saunier, N., Włodarek, P., Horst van der, R., Daniels, S. (2017). *Review of current study methods for VRU safety, Appendix 6 – Scoping review: Surrogate measures of safety in site-based road traffic observations*. Warsaw University of Technology, Poland

Lewis-Evans, B. (2004). *Traffic safety.* Michigan: Science Serving Society of Bloomfield Hills.

Lewis-Evans, B., & Charlton, S. G. (2006). Explicit and implicit processes in behavioural adaptation to road width. *Accident Analysis & Prevention, 38*, 610-617.

Lewis-Evans, B., & Rothengatter, T. (2009) Task difficulty, risk, effort and comfort in a simulated driving task – Implications for Risk Allostasis Theory. *Accident Analysis and Prevention, 41*, 1053-1063.

Lincoln, N. B., Taylor, J. L., Vella, K., Bouman, W. P., & Radford, K. A. (2010). A prospective study of cognitive tests to predict performance on a standardised road test in people with dementia. *International Journal of Geriatric Psychiatry*, *25*, 489-496.

Machin, A. M., & Sankey, K. S. (2008). Relationships between young drivers' personality characteristics, risk perceptions, and driving behavior. *Accident Analysis & Prevention*, *4,* 541-547.

McCormick, E. J., & Tiffin, J. (1980). *Industrial psychology*. Englewood Cliffs, N.J.: Prentice-Hall.

Münsterberg, H. (1913). *Psychology and industrial efficiency.* Boston, MA: Houghton Mifflin.

Nordfjærn, T., & Rundmo, T. (2013). Road traffic safety beliefs and driver behaviour among personality subtypes of drivers in the Norwegian population. *Traffic Injury Prevention*, *14*, 690-696.

Oltedal, S., & Rundmo, T. (2006). The effects of personality and gender on risky driving behaviour and accident involvement. *Safety Science*, *44*, 621-628.

Ozkan T., Lajunen T., & Summala, H. (2006). Driver Behaviour Questionnaire: A follow-up study. *Accident Analysis & Prevention, 38*, 386-395.

Raven, J., Raven, J. C., & Court, J. H. (2000). *Raven Manual: Section 3. SPM Manual (Including the Parallel and Plus Versions)*. Oxford: Oxford Psychologists Press.

Risser, R., Chaloupka, Ch., Grundler, W., Sommer, M., Hausler, J., & Kaufmann, C. (2008). Using non-linear methods to investigate the criterion validity of traffic-psychological test batteries. *Accident Analysis and Prevention, 4,* 149-157.

Sartori, R., & Pasini, M. (2007). Quality and quantity in test validity: How can we be sure that psychological tests measure what they have to? *Quality & Quantity, 41*, 359-374.

Schaie, K. W. (1994). The course of adult intellectual development. *American Psychologist*, *49*, 304-314.

Shannon, H. S., Robson, L. S., & Guastello, S. J. (1999). Methodological criteria for evaluating occupational safety intervention research. *Safety Science, 31,* 161-179.

Sommer, M., Herle, M., Hausler, J., Risser, R., Schutzhofer, B., & Chaloupka, Ch. (2008). Cognitive and personality determinants of fitness to drive. *Transportation Research Part F: Traffic Psychology and Behaviour*, *11*, 362-375.

Summala, H. (2000). Brake reaction times and driver behavior analysis. *Transportation Human Factors, 2*, 217-226.

Sundström, A. (2008). Self-assessment of driving skill: A review from a measurement perspective. *Transportation Research Part F: Traffic Psychology and Behaviour*, *11*(1), 1-9.

Svensson, Å., & Hydén, C. (2006). Estimating the severity of safety related behaviour. *Accident Analysis & Prevention, 38,* 379-385. http://dx.doi.org/10.1016/j.aap.2005.10.009

Szalma, J. L. (2009). Individual differences in human-technology interaction: Incorporating variation in human characteristics into human factors research and design. *Theoretical Issues in Ergonomics Science*, *1*, 381-397.

Thompson, K. R., Johnson, A. M., Emerson, J. L., Dawson, J. D., Boer, E. R., & Rizzo, M. (2012). Distracted driving in elderly and middle-aged drivers. *Accident Analysis & Prevention, 45,* 711-717.

Ulleberg, P., & Rundmo, T. (2003). Personality, attitudes and risk perception as predictors of risky driving behaviour among young drivers. *Safety Science 41*, 427-443.

Wontorczyk, A. (2011). *Niebezpieczne zachowanie kierowców. Psychologiczny model regulacji zachowań w ruchu drogowym*. Kraków: Wydawnictwo UJ.

Zhang, G., Yau, K. K., & Chen, G. (2013). Risk factors associated with traffic violations and accident severity in China. *Accident Analysis & Prevention*, *59*, 18-25.