Accident Analysis and Prevention 128 (2019) 8–16

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Assessing the relationship between self-reported driving behaviors and driver risk using a naturalistic driving study

Xuesong Wang[a,b,c,*], Xiaoyan Xu[a,c]

[a] College of Transportation Engineering, Tongji University, Shanghai, 201804, China
[b] The Key Laboratory of Road and Traffic Engineering, Ministry of Education, Shanghai, 201804, China
[c] National Engineering Laboratory for Integrated Optimization of Road Traffic and Safety Analysis Technologies, 88 Qianrong Rd, Wuxi 214151, China

ARTICLE INFO

Keywords:
Driver risk
Naturalistic driving study
Driver Behavior Questionnaire
Crash and near crash
Cluster analysis
Logistic regression

ABSTRACT

The Manchester Driver Behavior Questionnaire (DBQ) identifies risky driving behaviors resulting from psychological mechanisms. Investigating the relationships between these behaviors and drivers' crash risk can provide a better understanding of the personal factors contributing to the incidence of crashes, allowing the more effective development of safety education and road management countermeasures and interventions. The objectives of this study are therefore: 1) to determine the extent to which driver involvement in both crashes and near crashes (CNCs) is related to self-reported driving behaviors, and 2) to assess the relationship between each type of risky behavior and individual driver CNC risk. Driver and crash data were acquired from the Shanghai Naturalistic Driving Study and included 45 males and 12 females, participants with the mean age of 38.7. A K-mean cluster method was adopted to classify participants into three CNC groups of high-, moderate- and low-risk drivers. Drivers completed the DBQ to self-evaluate the frequency during their daily driving of the questionnaire's 24 risky behaviors. Principal component analysis of the 24 items led to a five-component structure including aggressive violations, ordinary violations, lapses, inattention errors, and inexperience errors. Two logistic regression models were developed to investigate the correlation between the five DBQ components and drivers' CNC levels. Conclusions are as follows: 1) high-risk drivers were significantly more likely to have engaged in inattention errors (e.g., missing a "yield" sign) and ordinary violations (e.g., running a red light) than the other drivers, and, 2) aggressive violations (e.g., racing against others) and ordinary violations were positively related to the probability of being a high- or moderate-risk driver.

1. Introduction

Human factors play a pivotal role in traffic crashes. Dingus et al. (2006) found that driver inattention was a contributing factor in 78% of the crashes in the United States' 100-car Study database, and a recent study by the American Automobile Association estimated that 56% of fatal crashes occurring between 2003 and 2007 involved aggressive driving behavior (AAA Foundation, 2009). In Shanghai, China, 792 of the 1044 police-reported crashes occurring in 2015 resulted from risky driving behaviors such as violations and inattention, accounting for 75.9% of total crashes (Shanghai Public Security Bureau, 2016). Evans and Wasielewski (1982) have shown that risky driver behaviors and crash involvement are associated at the individual level.

As behavior is an individual's external response to his or her mental activity, it has been demonstrated that various risky driving behaviors can result from particular attitudes and psychological traits. Owsley et al. (2003) have reported that people with higher impulsivity and empathy levels were prone to making more driving errors. Drivers with higher aggressiveness levels were shown to be more easily irritated (Stephens and Groeger, 2009), and tended to commit more violations (King and Parker, 2008). Determining which behaviors are most related to crash risk can help researchers pinpoint the underlying psychological traits that are, themselves, most related to crashes, in order to better identify potentially high-risk drivers. A more thorough understanding of the human factors that are associated with high-risk drivers can lead to the development of more effective safety countermeasures, such as improved driver training and testing, education campaigns aimed at changing driving practices, legislation to regulate driver behavior, and improvements to the design of road systems and automobiles (Elander et al., 1992), all of which have been found to decrease risky behaviors (Abele and Møller, 2011; Mccartt et al., 2003; Wang et al., 2018) and may consequently reduce an individual's involvement in crashes.

[*] Corresponding author at: College of Transportation Engineering, Tongji University, Shanghai, 201804, China.
E-mail address: wangxs@tongji.edu.cn (X. Wang).

https://doi.org/10.1016/j.aap.2019.03.009
Received 28 November 2018; Received in revised form 15 March 2019; Accepted 19 March 2019
Available online 05 April 2019

One measure of risky behaviors and bad habits in daily driving is the self-reported Manchester Driver Behavior Questionnaire (DBQ). The 24-item modified version used in our analysis is derived from the original 50-item questionnaire, and retains the eight highest-load items on each of the DBQ's three subscales: errors, lapses and violations (Parker et al., 1995a, 1995b). Errors are defined as misjudgments or failures of observation that could be hazardous to others, such as not seeing a pedestrian crossing; lapses are unintentional behaviors performed because of inattention or deficits such as taking the wrong exit; and violations are deliberate deviations of legally regulated or socially accepted behaviors associated with safe vehicle operation, such as speeding or close following of another vehicle (Zhao et al., 2012).

The three DBQ subscales of errors, lapses and violations are highly correlated with each other, however, so using them directly in regression analysis can lead to multicollinearity issues and biased inference. Principal component analysis (PCA), a prevalent method of solving such problem, is used, therefore, to divide the 24 DBQ items into several uncorrelated components. Three-component and four-component structures are the most common in existing studies. Utilizing a sample of Australian drivers, for example, Blockley and Harthy (1995) divided the DBQ items into three components: general errors, dangerous errors and dangerous violations. Aberg and Rimmo (1998) divided the items into four components: inattention errors, inexperience errors, lapses and violations. Based on Iranian drivers who overtook adjacent vehicles on rural roads, Kashani et al. (2016) obtained a four-component structure: errors, lapses, ordinary violations, and aggressive violations, the same four-component structure earlier produced by a Finnish sample (Mesken et al., 2002).

Using the DBQ to identify specific risky behaviors has at least three advantages for this study. First, the DBQ design is based on the idea that the three subscales (violations, errors and lapses) differ in their psychological mechanisms, and therefore in the kinds of remedial actions necessary to combat them (Reason et al., 1990). That is, exploring the association between the DBQ subscales and crash involvement can help researchers determine the likelihood of crash involvement for each type of risky behaviors and develop appropriate countermeasures. Second, the DBQ's detailed depictions of risky behaviors permits the driver's self-assessment that is important to identifying the underlying psychological mechanisms. Although self-reported data has obvious drawbacks, the self-assessment makes visible the behaviors otherwise difficult for researchers to observe and evaluate: compared with objectively measured risky driving behaviors such as unusual vehicle kinematics (e.g., high deceleration, low time-to-collision), DBQ items concentrate on the drivers' own perceptions and attitudes. The third advantage is that the DBQ's validity has been demonstrated, that is, safety-relevant links have been found to exist between DBQ scores and objectively measured risky driving behaviors (Zhao et al., 2012).

Significant relationships between DBQ scores and self-reported crash involvement have also been shown by previous research. Violations scores, in particular, have often been reported to be positively related to crash rates (Parker et al., 1995a, 1995b). Parker and West (1995) expanded on this study by dividing crashes into two categories, active and passive, based on the reporting driver's role in the crashes. The results indicated that both active and passive crash involvements were associated with violation scores. Additionally, a more recent study indicated that the combination of error and lapse scores were predictive of crashes (Afwahlberg et al., 2011). A meta-analysis of research using DBQ indicated that, of 76 datasets that met the criteria for analysis, 42 showed significant correlation between crash rates and violation scores, and 32 showed significant correlation between crash involvement and error scores (Dewinter and Dodou, 2010).

However, self-reported crashes can contain errors and omissions, and even accurately reported, the relatively low incidence of crashes in real road driving means that individual difference is rarely large enough to allow researchers to reach statistically significant conclusions that can help identify potentially high-risk drivers. Martinussen et al. (2017) recently conducted a unique large-sample study (N = 3683) of the relationship between the DBQ and recorded crashes. Even with this sample size, only 1.1% of participants were involved in a crash; and the low, moderate, and high-risk driver groups, as identified by their DBQ violation scores, did not differ in recorded crashes.

Traffic experts solve the low crash incidence problem by employing crash surrogates. A commonly used crash surrogate for driver safety assessment (Guo and Fang, 2013; Wu et al., 2014) is the near crash, defined as "any circumstance that requires a rapid evasive maneuver by the subject vehicle to avoid a crash" (Dingus et al., 2006). Near crashes meet the two principles of crash surrogate measures: a) they share the same or similar causal mechanism with crashes; and b) they have a strong frequency relationship with crashes. Near crashes also have similar kinematic signatures (Dingus et al., 2006), and Guo et al. (2010) has demonstrated their effectiveness in safety assessment. Sensitivity analysis has shown that combining crashes and near crashes (CNCs) can increase statistical power when identifying significant risk factors, thus providing more precise estimations of risk factors (Guo and Fang, 2013; Mcgehee et al., 2007).

One means of collecting CNC data is the naturalistic driving study (NDS). In NDS, vehicles are instrumented with an on-board data acquisition system (DAS) that continuously records video data and vehicle kinematics during normal daily driving. The rich video data collected in DAS makes it possible to extract near crashes as well as crashes. Additionally, the video cameras enable the recording of detailed information regarding road environment, vehicle movement, driver behavior and other factors that may contribute to a CNC. Using NDS data thus makes two important improvements over self-reported crashes: 1) by using near crashes as well as crashes, the sample size is large enough to support statistical analysis on an individual driver level; and 2) unlike self-reported crashes, the detailed data extracted for each CNC is video-recorded and can be repeatedly verified and analyzed.

Previous related research can be categorized into two types: assessing the relationship between 1) CNC risk and objectively measured behaviors such as abnormal deceleration and acceleration (Ashley et al., 2017; Guo and Fang, 2013), and between 2) crashes (not including near-crash) and DBQ self-reported behaviors (Parker et al., 1995a, 1995b; Reason et al., 1990). Each type of research has limitations. As mentioned, assessment by objectively measured behaviors overlooks driver attitudes and other psychological mechanisms that may be present before they give rise to the objective behaviors; while the crash-only studies are limited by the potential inaccuracies of self-reported crashes, and by the low incidence of crashes under actual driving conditions. This study aims to fill gaps in both assessment categories by, in the first case, examining the relationship between CNC risk and driver attitudes. In the second case, it assesses DBQ behaviors in relation to 1) naturalistic driving study data, and 2) CNC, that is, it expands the event data by including involvement in near crashes as well as crashes. To summarize, this study: 1) extracts CNC from the Shanghai Naturalistic Driving Study (SH-NDS) data; 2) classifies the drivers into risk groups based on their involvement in CNC; 3) convert the 24-item DBQ into several uncorrelated components that are usable for regression analysis; and 4) establishes a regression model to identify the risky driving behaviors that are correlated to drivers' risk levels.

2. Material and method

2.1. Extraction of crashes and near crashes

The Shanghai Naturalistic Driving Study (SH-NDS), the first naturalistic driving study in China, was conducted from December 2012 to December 2015 with the collaboration of Tongji University, General Motors, and the Virginia Tech Transportation Institute. Fifty-seven drivers (with ages ranging from 25 to 59, and driving experience ranging from 1 to 23 years) participated in the study, including 45 males and 12 females. The participants' vehicles were equipped with advanced data acquisition systems which include four camera views (forward, face, hand, rear), GPS, speedometer, three-dimension accelerometer, and radar. The three-year study collected data for 19,133 trips and 161,055 km in total.

While it is obvious whether or not a crash has occurred, the detection of near-crashes requires the identification of subtler signals, and the rich kinematic and video data collected in the SH-NDS makes the CNC extraction more efficient. Like crashes, near crashes are typically characterized by unusual vehicle kinematics such as high lateral and longitudinal acceleration and low time-to-collision (TTC).

Kinematic triggers were utilized in this study to identify possible events from the raw data. Thresholds followed the initial trigger criteria set in the 100-car Naturalistic Driving Study (Dingus et al., 2006; Bagdadi, 2013):

- Trigger 1: lateral motion equal to or greater than 0.7 g (gravitational acceleration);
- Trigger 2: longitudinal acceleration or deceleration equal to or greater than 0.6 g;
- Trigger 3: event button activated by the driver pressing a dashboard button when an event occurred that he/she deemed critical;
- Trigger 4: lateral motion equal to or greater than 0.5 g coupled with a forward TTC of 4 s or less;
- Trigger 5: longitudinal acceleration or deceleration equal to or greater than 0.5 g coupled with a forward TTC of 4 s or less.
- In order to ensure a sufficient sample size and the validity of extracted CNCs, the extraction process was designed as shown in Fig. 1.

It should be noted that: 1) accepting a fairly high false alarm rate (90%) ensured that few valid events were missed; and 2) collected events were screened by data reductionists watching forward and face video for a total of 20 s for each event, that is, from 10 s prior to and 10 s after the event. Events were manually validated as CNC if at least two of the following were observed: 1) a rapid, evasive maneuver by the subject vehicle during the event; 2) a conflict with another object or traffic participant, as seen from the front camera video; 3) a change in the subject driver's facial expression. Events were retained if verified as valid CNCs and discarded if not.

In order to produce an adequate sample size, changes were made to the initial CNC extraction trigger criteria, which are summarized in Table 1.

When using the initial criteria, the false alarm rates for Triggers 1, 3 and 4 were acceptable, that is, over 90%. In consideration of the labor intensity of manual validation, these two trigger types were therefore not adjusted. It can be seen in Table 1, however, that after adjusting the criteria for Triggers 2 and 5 to more liberal levels, the sample size increased substantially. Through the combination of trigger criteria and manual validation, 583 near crashes and 8 crashes were finally identified.

2.2. Factor analysis of the driver behavior questionnaire

All 57 participating drivers completed the modified Manchester Driver Behavior Questionnaire just before the SH-NDS period, which consists of 24 items describing scenarios related to risky behaviors or bad habits during daily driving. Based on their driving over the past year, drivers were required to judge the frequency of each item on a 6-point Likert scale (1 = never to 6 = almost every time). The mean and the standard deviation for each item are shown in Table 2.

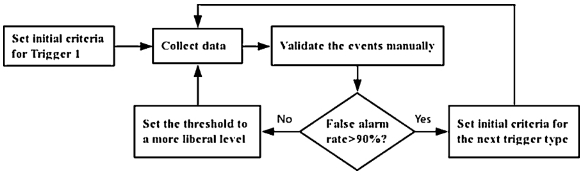

Set initial criteria for Trigger 1
Collect data
Validate the events manually
Set the threshold to a more liberal level
No
False alarm rate>90%?
Yes
Set initial criteria for the next trigger type

Fig. 1. CNC extraction process.

It can be seen in Table 2 that Item 15 elicited the same "never" response from all 57 drivers (mean of 1, standard deviation of 0). Hence, we deleted Item 15 from the DBQ when conducting subsequent analysis.

Because the 23 remaining DBQ items were highly correlated with each other, including all the items in a regression model would have led to multicollinearity issues and biased inference. In order to resolve the multicollinearity and retain the maximum information from the questionnaire, principal component analysis (PCA) was performed. PCA uses an orthogonal transformation to convert correlated variables into a set of uncorrelated variables called principal components (Jolliffe, 2011). To determine suitability for component analysis, the Kaiser-Meyer-Olkin Test and Bartlett's test were conducted. The results, which respectively, were generally satisfactory (KMO = 0.654) and showed great significance (P-value < 0.001), confirmed suitability. The eigenvalues for each principle component were used as main indicators to select significant components. The eigenvalues, ordered by size, are shown in Table 3.

As can be seen from Table 3, PCA results showed seven components with eigenvalues greater than 1, which meant these seven components contained substantial information. The curve in the scree plot (Fig. 2), however, flattens after the fifth component, which indicates the first five components contributed most to the total variance. Thus, a five-component structure was adopted, accounting for a cumulative 58.219% of the total variance.

In order to define the 5 significant components, we constructed the component score coefficient matrix shown in Table 4. Because questionnaire items with high loadings are of greater importance, the items with coefficients greater than 0.2 for each component are marked in red bold. The table is ordered by the highest loadings for each component, and items with loadings less than 0.20 in any component are omitted.

Each questionnaire item is described with its original type, or subscale (V, E, or L), per Parker and West (1995). Note that within all five components, the types for all listed items, those with high loadings, conform with each other.

Components 1 and 5 both contain violation items. Items in Component 1, however, represent greater aggressiveness towards other drivers than those in Component 5. Thus, Components 1 and 5 were defined as aggressive violations and ordinary violations respectively. Likewise, Components 3 and 4 both contain error items. The errors in Components 3 and 4 are caused, respectively, by inattention (e.g., inattention to pedestrians and cyclists) and inexperience (e.g., brake too quickly on a slippery road); therefore, these two components were distinguished as inattention errors and inexperience errors.

In accord with the above analysis, the five components were finally defined as follows:

- Component 1: aggressive violations (AV)
- Component 2: lapses (L)
- Component 3: inattention errors (IAE)
- Component 4: inexperience errors (IEE)
- Component 5: ordinary violations (OV)

Each component can be expressed by a linear combination of its high-load items. The combinations allowed each of the five components to be calculated for each driver, using R Project. The greater a driver's component value, the higher is his or her frequency of the risk behaviors represented by that component. The calculation results for all drivers are shown in Fig. 3, in which, each histogram consists of 30 contiguous bars. Each bar occupies an equal interval length on the X-axis, which measures the component value. Each bar's height represents the count of drivers whose value is within that interval.

As can be seen in Fig. 3, each component obeys a right-skewed

Table 1
Changes in CNC extraction criteria by trigger type.

Trigger Type	Initial Criterion	Final Criterion	Increase of Sample Size
1. Lateral motion	$\geq 0.7g$	$\geq 0.7g$	0
2. Longitudinal acceleration or deceleration	$\geq 0.6g$	$\geq 0.5g$	281%
3. Event button	Activated	Activated	0
4. Lateral motion & forward time-to-collision	Lateral motion$\geq 0.5g$; and $TTC \leq 4s$	Lateral motion$\geq 0.5g$; and $TTC \leq 4s$	0
5. Longitudinal acceleration or deceleration & forward time-to-collision	longitudinal acceleration $\geq 0.5g$ or deceleration $\leq -0.5g$; and $TTC \leq 4s$	longitudinal acceleration $\geq 0.45g$ or deceleration $\leq -0.45g$; and $TTC \leq 4s$	207%

Table 2
Means and standard deviations of the 24 DBQ items.

No	Item	Mean	S.D.
1	Attempt to drive away from traffic lights in the wrong gear	1.183	0.469
2	Become impatient with a slow driver in the fast lane and pass on the forbidden side	2.267	1.191
3	Drive especially close to a car in front as a signal to the driver to go faster or get out of the way	1.600	0.924
4	Attempt to pass someone that you hadn't noticed to be making an offside turn	1.467	0.812
5	Forget where you left your car in a parking lot	2.150	1.005
6	Turn on one instrument, such as your headlights, when you mean to switch on something else, such as the windshield wipers	1.483	0.651
7	Realize that you have no clear recollection of the road along which you have just been traveling	1.850	0.685
8	Cross an intersection knowing that the traffic lights have already changed from yellow to red	1.633	0.882
9	Fail to notice that pedestrians are crossing when turning onto a side street from a main road	1.250	0.437
10	Angered by another driver's behavior, you catch up to them with the intention of giving him/her "a piece of your mind."	1.333	0.655
11	Misread the signs and turn the wrong direction on a one-way street	1.567	0.621
12	Disregard the speed limits late at night or early in the morning	1.883	1.010
13	When making a near side turn, nearly hit a bicyclist (or moped rider) who is riding along side of you	1.400	0.588
14	Attempting to turn onto a main road, you pay such close attention to traffic on the road entering that you nearly hit the car in front of you that is also waiting to turn	1.317	0.504
15	Drive even though you realize you might be over the legal blood alcohol limit	1.000	0.000
16	Have an aversion to a particular class of road user, and indicate your hostility by whatever means you can	1.583	0.787
17	Underestimate the speed of an oncoming vehicle when attempting to pass a vehicle in your own lane	1.383	0.555
18	Hit something when backing up that you had not previously seen	1.550	0.534
19	Intending to drive to destination A, you "wake up" to find yourself on a road to destination B, perhaps because destination B is a more common destination	1.450	0.594
20	Get into the wrong lane approaching an intersection	1.950	0.790
21	Miss "yield" sign, and narrowly avoid colliding with traffic having the right of way	1.267	0.482
22	Fail to check your rearview mirror before pulling out, changing lanes, etc.	1.233	0.427
23	Get involved in unofficial "races" with other drivers	1.283	0.666
24	Brake too quickly on a slippery road or steer the wrong way into a skid	1.233	0.427

Table 3
Eigenvalues for principle components.

Component	Eigenvalue	Proportion	Cumulative
1	6.285	27.325	27.325
2	2.581	11.221	38.546
3	1.866	8.112	46.658
4	1.428	6.208	52.866
5	1.231	5.353	58.219
6	1.201	5.223	63.441
7	1.156	5.024	68.466
8	0.953	4.145	72.611
…	…	…	…
23	0.089	0.386	100.000

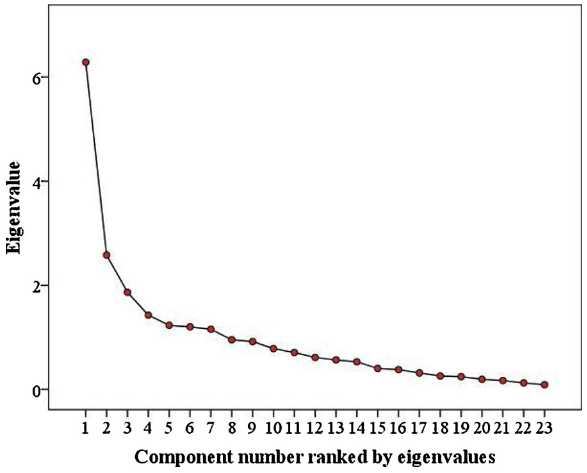


Fig. 2. PCA scree plot.

distribution in which the longer tail is on the right side, indicating that most drivers are concentrated in the low value region, and the higher the value, the smaller the number of drivers. The distribution confirms the expectation that the majority of drivers have a low frequency of any specific risky behavior or habit.

3. Methodology

3.1. Cluster analysis

Previous research relevant to the present study has widely adopted cluster analysis to investigate the differences in characteristics among driver groups, and has revealed that the cluster analysis of dependent variables can produce robust regression results (Guo and Fang, 2013; Donmez et al., 2010). In this study, each driver's CNC rate was calculated to represent his or her CNC involvement:

$$CNC\ Rate_i = \frac{N_i}{M_i} \tag{1}$$

where $CNC\ Rate_i$ is the number of CNCs per 100 km for individual i; N_i is the total number of CNCs for individual i during the SH-NDS period;

Table 4
Component score coefficient matrix.

DBQ#	Item with Original Type (Subscale)	Components				
		1	2	3	4	5
23	Get involved in unofficial “races” with other drivers (V)	0.310	−0.070	0.007	0.044	−0.121
16	Have an aversion to a particular class of road user, and indicate your hostility by whatever means you can (V)	0.293	0.146	0.037	−0.151	−0.196
12	Disregard the speed limits late at night or early in the morning (V)	0.276	−0.029	−0.037	0.060	−0.050
2	Become impatient with a slow driver in the fast lane and pass on the forbidden side (V)	0.227	−0.054	−0.232	0.104	0.194
1	Attempt to drive away from traffic lights in the wrong gear (L)	−0.004	0.363	−0.052	0.006	−0.133
19	Intending to drive to destination A, you “wake up” to find yourself on a road to destination B, perhaps because destination B is a more common destination (L)	−0.064	0.361	−0.040	−0.108	−0.081
20	Get into the wrong lane approaching an intersection (L)	0.040	0.255	−0.117	−0.041	0.165
9	Fail to notice that pedestrians are crossing when turning onto a side street from a main road (E)	−0.063	−0.103	0.443	−0.114	0.007
21	Miss “yield” sign, and narrowly avoid colliding with traffic having the right of way (E)	−0.030	−0.038	0.335	−0.059	−0.083
13	When making a near side turn, nearly hit a bicyclist (or moped rider) who is riding along side of you (E)	0.079	0.119	0.242	−0.097	−0.147
24	Brake too quickly on a slippery road or steer the wrong way into a skid (E)	0.010	−0.065	−0.019	0.447	−0.143
4	Attempt to pass someone that you hadn’t noticed to be making an offside turn (E)	−0.088	−0.033	−0.211	0.388	0.135
22	Fail to check your rearview mirror before pulling out, changing lanes, etc. (E)	0.008	−0.041	0.150	0.217	−0.091
17	Underestimate the speed of an oncoming vehicle when attempting to pass a vehicle in your own lane (E)	0.070	−0.157	0.159	0.213	0.148
3	Drive especially close to a car in front as a signal to the driver to go faster or get out of the way (V)	−0.114	0.027	−0.047	−0.011	0.474
8	Cross an intersection knowing that the traffic lights have already changed from yellow to red (V)	0.082	−0.222	0.150	−0.241	0.334
Variance explained		27.325	11.221	8.112	6.208	5.353

Note: V = violations, E = errors, L = lapses.

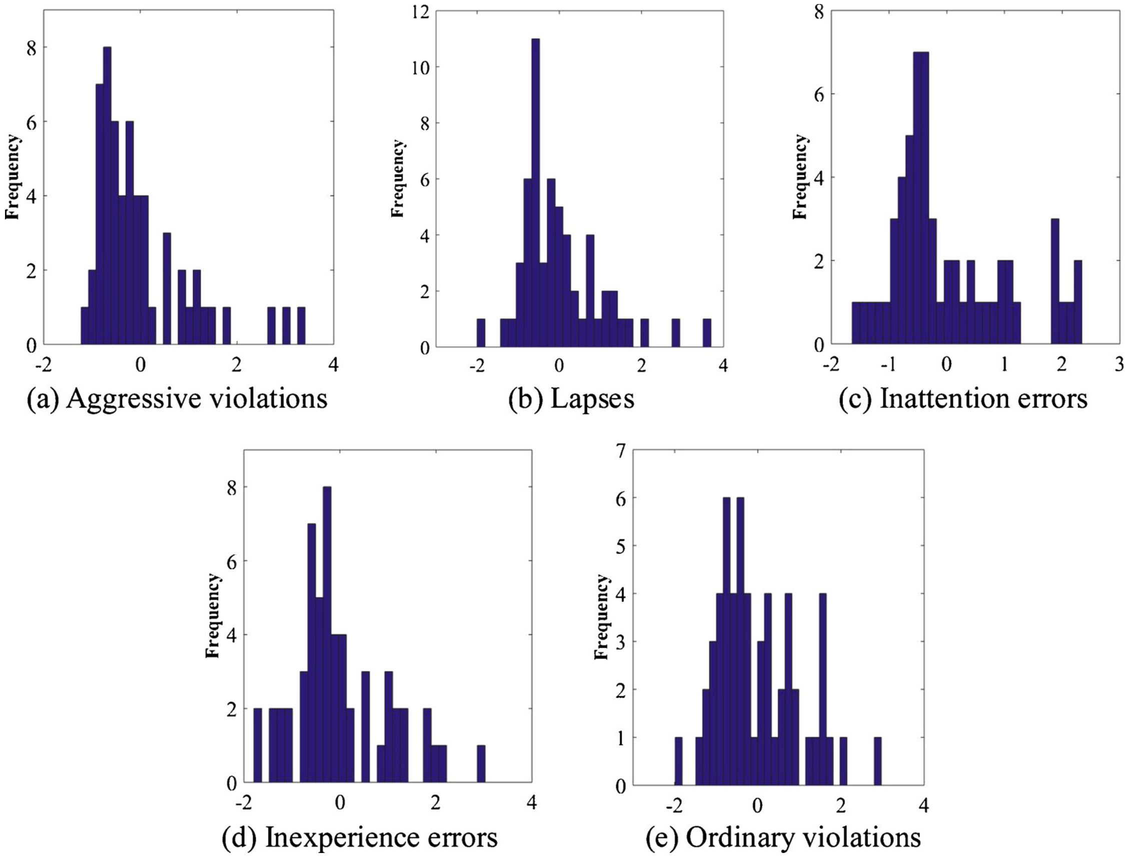
and M_i is the kilometers travelled by individual i during the period.

K-mean cluster analysis, which possesses several advantages such as algorithmic simplicity, fast calculation speed, and good clustering effect, was adopted to partition the observations into k clusters in which each observation belongs to the cluster with the nearest mean. The K-mean cluster method was utilized to classify the 57 drivers into 3 risk groups, high, moderate, and low, based on their CNC rates. In order to evaluate the clustering effect, the within-cluster sum of squares and average silhouette were considered. Within-cluster sum of squares is calculated as follows:

$$WCSS = \sum_{i=1}^{k} \sum_{X_j \in S_i} \|X_j - \mu_i\|^2 \tag{2}$$

where $(X_1, X_2, \ldots, X_n)$ is a set of observations in which each observation is a one-dimensional real vector representing, in this study, the CNC rate; $(S_1, S_2, \ldots, S_k)$ is the set of k clusters; and μ_i is the mean number of observations in set S_i.

The average silhouette is another widely used criterion for



(a) Aggressive violations (b) Lapses (c) Inattention errors

(d) Inexperience errors (e) Ordinary violations

Fig. 3. Distribution of the five components among all drivers.

evaluating the clustering effect. The silhouette of an observation measures how closely it is matched to other observations within its cluster and how loosely it is matched to the observations in neighboring clusters (Rousseeuw, 1987). The equation is as follows:

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \tag{3}$$

where $s(i)$ is the silhouette of observation i; $a(i)$ is the average distance between i and all other observations within the same cluster; and $b(i)$ is the average distance of i to the observations in neighboring clusters. $s(i)$ range from -1 to $+1$. The closer $s(i)$ is to $+1$, the more appropriately i is clustered. The average silhouette over all observations of the entire dataset is a measure of how appropriately the data have been clustered.

3.2. Logistic regression analysis

After the drivers were partitioned into risk groups, a logistic regression model was established to estimate the probability of each individual being a high-risk driver.

The dependent variable for individual i is the binary variable y_i, which is assumed to follow a Bernoulli distribution. The variable is defined as:

$$y_i = \begin{cases} 1\, if\ driver\ i\ is\ in\ the\ high-risk\ group \\ 0\, otherwise \end{cases} \tag{4}$$

p_i is defined as the possibility of y_i being 1. Thus, the logistic regression model is as follows:

$$p_i = p(y_i = 1X_i) = \frac{exp(\alpha + \beta X_i)}{1 + exp(\alpha + \beta X_i)} \tag{5}$$

where X_i is the matrix of predictors for individual i; and α and β are the vectors of regression parameters. In this study, the predictors were the five components extracted from DBQ by PCA. Based on the predictors, the model estimated the probability of a driver being at high CNC risk. A driver was predicted as high-risk if this probability was greater than a predefined threshold value p_0 which is usually set to 0.5.

The predictive ability of the logistic regression model can be illustrated by a receiver operating characteristic (ROC) curve, which plots the true positive rate (also known as sensitivity) against the false negative rate (1-specificity) with p_0 varying from 0 to 1. In this study's logistic regression model, sensitivity is the probability of correctly predicting y_i as 1, that is, correctly predicting driver i is in a high-risk group; specificity is the probability of correctly predicting driver i is not in the group. The predictive performance can be measured by the area under the curve (AUC), which varies between 0 and 1. The higher the AUC value, the better the prediction power generated by the model.

4. Results

4.1. Results of the cluster analysis

As noted above, to aid interpretation, K-mean clusters were used to classify the 57 drivers into three predetermined risk groups, to represent high, moderate and low risk. Under the three-group cluster scheme, the total within-cluster sum of squares and average silhouette equaled 0.28 and 0.59, which demonstrated a fairly good clustering effect. The output of cluster analysis based on CNC rate is shown in Fig. 4.

As can be seen from Fig. 4, the large low-risk group included 7 drivers that cannot be shown in the figure due to their CNC rates of zero. Just over half of all drivers were classified into the higher risk groups: 8 drivers in the high-risk group and 23 in the moderate-risk group. Detailed characteristics of the three groups are summarized in Table 5.

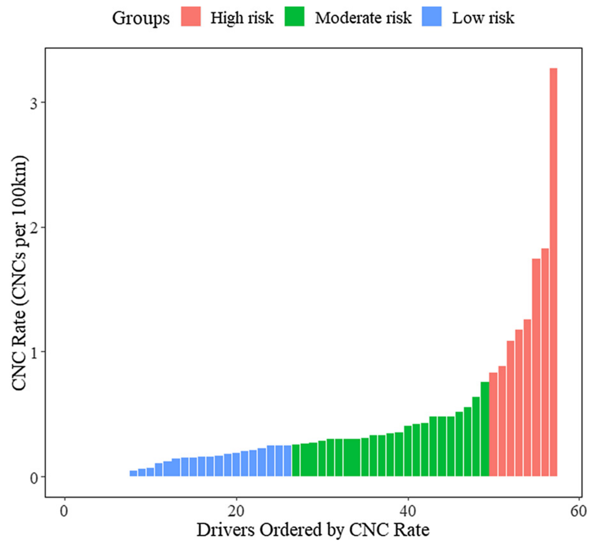


Fig. 4. Distribution of CNC rates by driver risk groups.

It can be noted in Table 5 that the mean age does not show great difference among the three risk groups. However, the percentage of male drivers is much lower in the high-risk group than in the other groups, and lower than their 78.95% representation in the overall sample (45 male to 12 female); while the percentage in the low-risk group is closer to the sample proportion, and their percentage is higher than sample in the moderate-risk group. The mean CNC rate of the high-risk group is almost 4 times that of the moderate-risk group, and 12.5 times that of the low-risk group. The overall pattern of the five DBQ components indicates that 1) the high-risk group had the highest values among the three groups in inattention errors and ordinary violations; 2) the moderate-risk group had the highest values in aggressive violations and inexperience errors, and also had the second highest value in ordinary violations.

The boxplots in Fig. 5 visually compare distributions of the five DBQ components among the three risk groups. Differences in inattention error and ordinary violation distributions are readily apparent.

4.2. Results of logistic regression models

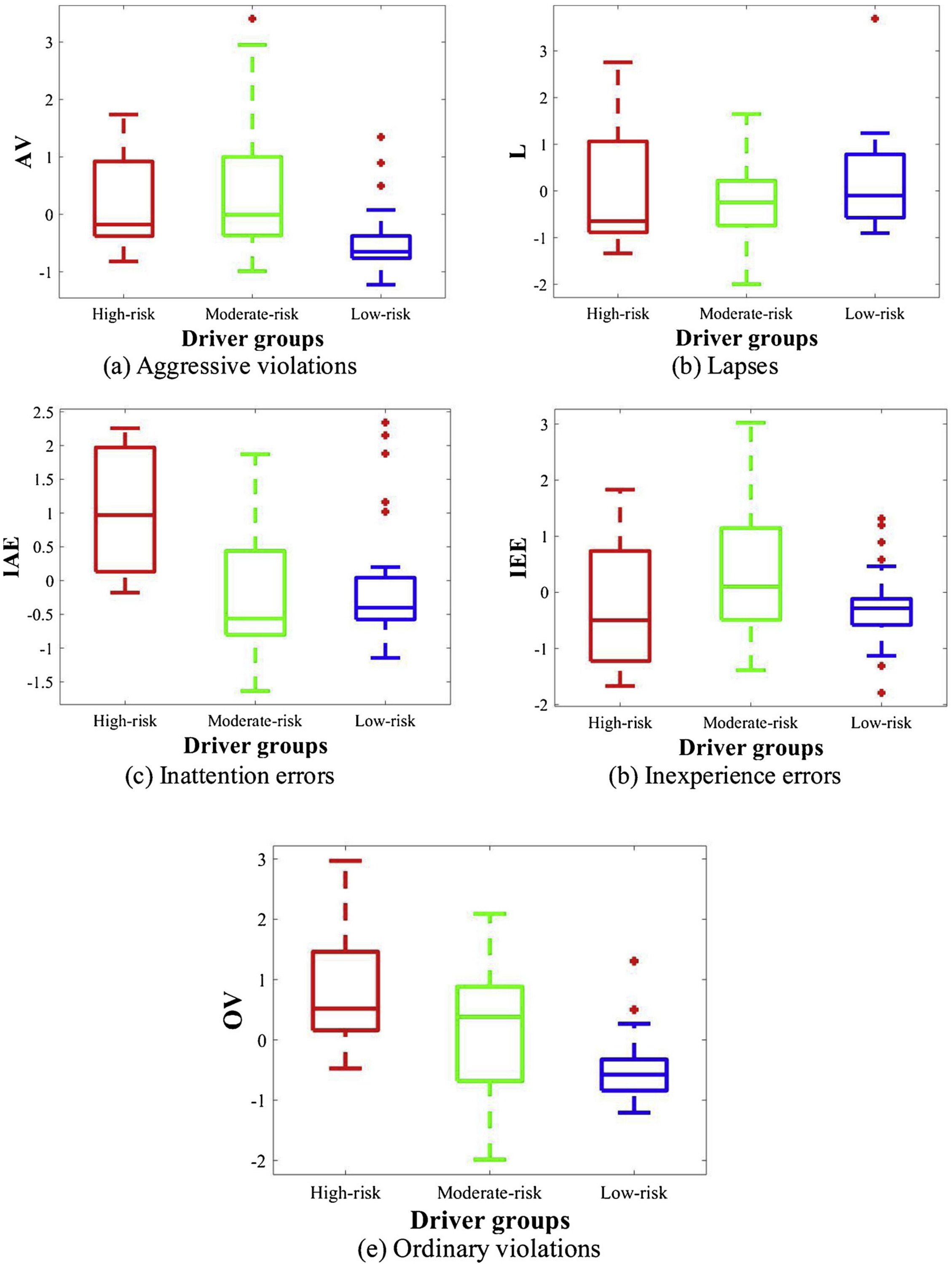
While the CNC-based cluster analysis identified drivers as belonging to the three risk groups, and the means of the five DBQ components gave an overall sense of the self-reported risky behaviors' distribution among the groups, the key question was in exactly which ways the components were related to the drivers' risk levels. As our interest is in drivers with extremely high and moderate to high CNC rates, two logistic regression models were developed. The first model emphasizes the comparison between high-risk drivers and moderate- to low-risk drivers, while the second compares high- and moderate-risk drivers to low-risk, or safe drivers. The outputs of the two models are summarized in Table 6.

In Model 1, inattention errors and ordinary violations had significant impact on the probability of being classified as a high-risk driver, with both the estimated parameters having positive values. In other words, drivers who frequently engaged in these types of behavior during daily driving were at the highest risk level for involvement in CNCs. The odds ratio (OR) represents the relative odds of being a high-risk driver for every one-unit increase in a continuous variable. Model 1′s OR of ordinary violations (= 3.85) therefore indicated that every one-unit increase in ordinary violations would result in an almost threefold (3.85 - 1 = 2.85) increase in the probability of high-risk classification. The OR for inattention errors (= 4.76) was even higher: every one-unit increase would multiply the probability of classification as a high-risk driver by almost 5.

In Model 2, aggressive violations and ordinary violations were the

Table 5
Characteristics of risk groups.

Risk group	Number of drivers	Mean age	Male percentage	Mean CNC rate	Means of the 5 DBQ components				
					AV	L	IAE	IEE	OV
High	8	37.25	50.00%	1.51	0.21	0.06	1.03	−0.23	0.85
Moderate	23	38.09	91.30%	0.40	0.42	−0.20	−0.30	0.35	0.21
Low	26	39.69	76.92%	0.12	−0.44	0.16	−0.05	−0.24	−0.45



(a) Aggressive violations

(b) Lapses

(c) Inattention errors

(b) Inexperience errors

(e) Ordinary violations

Fig. 5. Distribution of the five DBQ components.

Table 6
Logistic regression models' outputs.

Models	Predictors	Parameter estimate	P-value	Odds ratio	95% odds ratio confidence limits	
Model 1: high vs. moderate/low risk	aggressive violations	0.61	0.20	1.85	0.72	4.75
	lapses	0.17	0.70	1.18	0.50	2.77
	inattention errors	**1.56**	**0.01***	**4.76**	**1.45**	**15.59**
	inexperience errors	−0.30	0.51	0.74	0.30	1.82
	ordinary violations	**1.35**	**0.03***	**3.85**	**1.13**	**13.11**
	intercept	−3.03	0.00	0.05		
Model 2: high/moderate vs. low risk	**aggressive violations**	**2.26**	**0.00****	**9.60**	**2.21**	**41.61**
	lapses	−0.97	0.06	0.38	0.14	1.04
	inattention errors	0.13	0.74	1.13	0.55	2.34
	inexperience errors	0.86	0.06	2.37	0.97	5.84
	ordinary violations	**1.87**	**0.00****	**6.47**	**2.05**	**20.46**
	intercept	0.90	0.06	2.47		

Note: boldfaced predictors have very low P-values: * means P-value < 0.05; ** means P-value < 0.01.

significant factors that positively correlated with the probability of being a high- or moderate-risk driver. That is, drivers who reported more aggressive and ordinary violations were at greater risk for CNC involvement than low-risk drivers. As with Model 1, ordinary violations continued to be a powerful predictor for drivers at risk for CNC. Model 2′s OR of ordinary violations (= 6.47) was much higher, in fact, than the OR for Model 1 (= 3.85), and indicated that, for every one-unit increase in ordinary violations, the relative odds of being classified as a higher risk driver would increase 5.47 times.

To evaluate model performance, receiver operating characteristics curves (ROCs) for Model 1 and Model 2 were plotted, and are illustrated in Fig. 6. The straight diagonal is a reference line for the colored ROC curves. The AUC value is 0.901 for Model 1 and 0.913 for Model 2. When compared with the perfect AUC value of 1, it is evident that the two models both generated outstanding predictive power.

5. Discussion and conclusions

To achieve the objective of assessing the relationship between self-reported risky behaviors and drivers' involvement in crashes and near crashes, data from 57 participants in the Shanghai Naturalistic Driving Study and their completed Manchester Driver Behavior Questionnaires were utilized for method development. A total of 591 CNCs extracted from the SH-NDS data were used to evaluate individual driver risk. Based on CNC rate, a K-mean cluster analysis indicated that 14% of drivers were at high risk for involvement in CNCs, and 40% drivers were at moderate risk, together representing over half of the participating drivers. The average CNC rate of the high-risk group was 3 times higher than the rate of the moderate-risk group, and 15 times higher than the rate of the low-risk group.

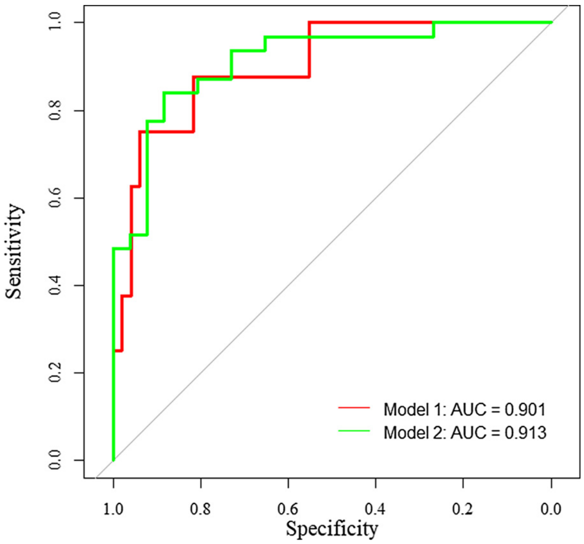


Fig. 6. Predictive performance of Model 1 (high vs. moderate/low) and Model 2 (high/moderate vs. low).

PCA was conducted to convert the DBQ's 24 correlated items into five uncorrelated components, which would effectively address the multicollinearity issues for regression analysis. According to the items with high loadings, the five components were defined as aggressive violations, ordinary violations, lapses, inattention errors, and inexperience errors. Compared with the three-subscale structure (violations, errors and lapses) originally developed by Parker and West (1995), this study made a further division of violations and errors to fit the PCA results. Similar solutions can be found in previous studies. As in this study, Aberg and Rimmo (1988) partitioned the errors into errors of inattention and inexperience, and Kashani et al. (2016) classified the violations into aggressive and ordinary violations.

The logistic regression models were developed to determine the extent to which driver CNC risk was related to self-reported driving behaviors. As illustrated in Table 6, Model 1 indicated that drivers who reported more inattention errors and ordinary violations were at particularly high CNC risk: every one-unit increase in inattention errors and ordinary violations, respectively, would multiply the probability of classification as a high-risk driver by almost 5 and 4. In Model 2, those who reported more aggressive violations and ordinary violations were more likely to be involved in CNCs (high and moderate risk) than safe, or low-risk drivers: every one-unit increase in aggressive violations and ordinary violations, respectively, would result in a nine-fold and six-fold increase in the probability of higher risk classification

This study's prediction of ordinary violations in both models conform with previous studies that have indicated that violations were the most powerful predictor of crash involvement (Reason et al., 1990; Parker et al., 1995a, 1995b). Referring back to Table 4, the DBQ ordinary violation component was composed of Items 3 and 8: 1) drive especially close to a car in front as a signal to the driver to go faster or get out of the way, and 2) cross an intersection knowing that the traffic lights have already changed from yellow to red. Both items clearly reflect the conclusions of previous researchers regarding the psychological mechanisms of violations, according to Parker et al. (1992), drivers who commit more violations were apt to underestimate the potentially negative consequences of such violations, perceive less social disapproval for the commission of violations, and find it difficult to voluntarily control such behaviors; Lucidi et al. (2010) pointed out these drivers tended to have high levels of normlessness and excitement-seeking. The high incidence of ordinary violations in this study, in conjunction with previous research on their underlying psychological mechanisms, suggests that CNC risk resulting from violations can be mitigated by confronting drivers' attitudes, beliefs, and norms with the goal of improving the safety culture overall.

This study also identified inattention error as another risky behavior highly related to CNC risk. The successful identification of inattention error may be owed to its separation from inexperience errors, a component that showed no significant association with CNC risk. The DBQ inattention errors include: 1) fail to notice that pedestrians are crossing when turning onto a side street from a main road; 2) miss “yield” sign, and narrowly avoid colliding with traffic having the right of way; 3) when turning nearside, nearly hit a bicyclist. Regan et al. (2011) defines inattention errors as the driver paying “insufficient or no attention to activities critical for safe driving brought by”: 1) failing to effectively distribute attention among multiple driving activities, 2) neglecting the activities critical for safe driving, 3) failing to fully complete the safety-critical activities in a hurry, and 4) diverting attention from the safety-critical driving activities to a secondary task (Regan et al., 2011). Consideration of the Regan et al. analysis suggests specific ways that CNC risk caused by inattention errors can be minimized by safety education. Such education could include, for example, practice in more appropriate attention prioritization, decision making, and response selection. Additionally, and perhaps more easily, advanced driver-assistance systems, such as the pedestrian collision warning system, can be applied more broadly to control inattention errors during daily driving.

Components that indicated nonsignificant associations with driver CNC risk included lapses as well as inexperience errors. Perhaps unexpectedly, high-risk drivers had the lowest mean value in inexperience errors among the three groups, and high- and moderate- risk drivers had lower mean values of lapses than low-risk drivers (Table 5). Since inexperience errors (IEE) and lapses (L) are both related to driving proficiency and experience, one possible reason for these results is erroneous self-reporting on the DBQ. That is, drivers with higher CNC risk may tend to be overconfident in their driving skills, resulting in the underestimation of their IEE and L frequency. The accuracy of self-reported inexperience errors and lapses would benefit from further validation by objective means.

Despite being limited by a relatively small sample size, this study succeeded in making a clear connection between self-reported risky driving behaviors and individual driver CNC risk. The results of this study can be used to improve safety education programs by increasing drivers' awareness of the behaviors that can lead to high CNC risk, and it can also provide a valuable reference for developing safety education regulations and other proactive safety countermeasures.

Acknowledgements

This study was jointly sponsored by the Chinese National Science Foundation (51878498), the Science and Technology Commission of Shanghai Municipality (18DZ1200200), the Chinese National Engineering Laboratory for Integrated Optimization of Road Traffic and Safety Analysis Technologies, and the Chinese 111 Project (B17032). The authors are grateful to Barbara Rau Kyle for her helpful edit and thoughtful feedback.

References

Abele, L., Møller, M., 2011. Relationship between road design and driving behavior: a simulator study. 3rd International Conference on Road Safety and Simulation.

Aberg, L., Rimmö, P.A., 1998. Dimensions of aberrant behaviour. Ergonomics 41 (1), 39–56.

Afwahlberg, A.E., Dorn, L., Kline, T., 2011. The Manchester driver behavior questionnaire as a predictor of road traffic accidents. Theor. Issues Ergon. Sci. 12 (1), 66–86.

American Automobile Association Foundation for Traffic Safety, 2009. Aggressive Driving: Research Update. American Automobile Association Foundation for Traffic Safety, Washington, D.C.

Ashley, G., Osman, O.A., Ishak, S., Codjoe, J., 2017. Effect of secondary tasks and driver behavior on crash/near-crash risk: naturalistic driving study. Presented at 96th Annual Meeting of the Transportation Research Board, Washington, D.C., 2017.

Bagdadi, O., 2013. Assessing safety critical braking events in naturalistic driving studies. Transp. Res. Part F Traffic Psychol. Behav. 16, 117–126.

Blockley, L., Harthy, L., 1995. Aberrant driving behavior: errors and violations. Ergonomics 38 (9) 1759-177l.

Dewinter, J.C.F., Dodou, D., 2010. The driver behaviour questionnaire as a predictor of accidents: a meta-analysis. J. Saf. Res. 41, 463–470.

Dingus, T.A., Klauer, S.G., Neale, V.L., Petersen, A., Lee, S.E., Sudweeks, J.D., Perez, M.A., Hankey, J.M., Ramsey, D., Gupta, S., Bucher, C., Doerzaph, Z.R., Jermeland, J., Knipling, R.R., 2006. The 100-car Naturalistic Driving Study: Phase II-Results of the 100-car Field Experiment. Report No.: DOT HS 810 593. National Highway Traffic Safety Administration, Washington, D.C.

Donmez, B., Boyle, L.N., Lee, J.D., 2010. Differences in off-road glances: effects on young drivers' performance. J. Transp. Eng. 136 (5), 403–409.

Elander, J., West, R., French, D., 1992. Behavioral correlates of individual differences in road traffic crash risk: an examination of methods and findings. Psychol. Bull. 113, 279–294.

Evans, L., Wasielewski, P., 1982. Do accident-involved drivers exhibit riskier everyday driving behavior? Accid. Anal. Prev. 14 (1), 57–64.

Guo, F., Fang, Y., 2013. Individual driver risk assessment using naturalistic driving data. Accid. Anal. Prev. 61, 3–9.

Guo, F., Klauer, S.G., Hankey, J.M., Dingus, T.A., 2010. Near crashes as crash surrogate for naturalistic driving studies. Transp. Res. Rec.: J. Transp. Res. Board 2147, 66–74.

Jolliffe, I.T., 2011. Principal component analysis. J. Mark. Res. 87 (100), 513.

Kashani, A.T., Ravasani, M.S., Ayazi, E., 2016. Analysis of drivers' behavior using Manchester Driver Behavior Questionnaire based on roadside interview in Iran. Int. J. Transp. Eng. 4 (1), 61–74.

King, Y., Parker, D., 2008. Driving violations, aggression and perceived consensus. Eur. Rev. Appl. Psychol. 58 (1), 43–49.

Lucidi, F., Giannini, A.M., Sgalla, R., Mallia, L., Devoto, A., Reichmann, S., 2010. Young novice driver subtypes: relationship to driving violations, errors and lapses. Accid. Anal. Prev. 42 (6), 1689–1696.

Martinussen, L.M., Møller, M., Prato, C.G., Haustein, S., 2017. How indicative is a self-reported driving behavior profile of police-registered traffic law offences? Accid. Anal. Prev. 99, 1–5.

Mccartt, A.T., Braver, E.R., Geary, L.L., 2003. Drivers' use of handheld cell phones before and after New York state's cell phone law. Prev. Med. 36 (5), 629–635.

Mcgehee, D.V., Raby, M., Carney, C., Lee, J.D., Reyes, M.L., 2007. Extending parental mentoring using an event-triggered video intervention in rural teen drivers. J. Safety Res. 38 (2), 215–227.

Mesken, J., Lajunen, T., Summala, H., 2002. Interpersonal violations, speeding violations and their relation to accident involvement in Finland. Ergonomics 45 (7), 469–483.

Owsley, C., McGwin, G.J., McNeal, S.F., 2003. Impact of impulsiveness, venturesomeness, and empathy on driving by older adults. J. Saf. Res. 34, 353–359.

Parker, D., Manstead, A.S.R., Stradling, S.G., Reason, J.T., 1992. Determinants of intention to commit driving violations. Accid. Anal. Prev. 24 (2), 117–131.

Parker, D., Reason, J.T., Manstead, A.S.R., Stradling, S.G., 1995a. Driving errors, driving violations and accident involvement. Ergonomics 38 (5), 1036–1048.

Parker, D., West, R., Stradling, S.G., Manstead, A.S.R., 1995b. Behavioral characteristics and involvement in different types of traffic accident. Accid. Anal. Prev. 27 (4), 571–581.

Reason, J.T., Manstead, A.S.R., Stradling, S.G., Baxter, J.S., Campbell, K., 1990. Errors and violations on the road: a real distinction? Ergonomics 33, 1315–1332.

Regan, M.A., Hallett, C., Gordon, C.P., 2011. Driver distraction and driver inattention: definition, relationship and taxonomy. Accid. Anal. Prev. 43 (5), 1771–1781.

Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20 (20), 53–65.

Shanghai Public Security Bureau Traffic Police Headquarters, 2016. Brief introduction of road traffic accidents in Shanghai. Traffic Transp. 2, 78–80.

Stephens, A.N., Groeger, J.A., 2009. Situational specificity of trait influences on drivers' evaluations and driving behavior. Transp. Res. Part F Traffic Psychol. Behav. 12 (1), 29–39.

Wang, X., Xing, Y., Luo, L., Yu, R., 2018. Evaluating the effectiveness of behavior-based safety education methods for commercial vehicle drivers. Accid. Anal. Prev. 117, 114–120.

Wu, K.F., Aguerovalverde, J., Jovanis, P.P., 2014. Using naturalistic driving data to explore the association between traffic safety-related events and crash risk at driver level. Accid. Anal. Prev. 72, 210–218.

Zhao, N., Mehler, B., Reimer, B., D'Ambrosio, L., Mehler, A., Coughlin, J., 2012. An investigation of the relationship between the Driver Behavior Questionnaire and objective measures of highway driving behavior. Transp. Res. Part F Traffic Psychol. Behav. 15 (6), 676–685.