IJTE, Vol. 14, No. 1, Summer 2026, pp. 2419-2434
Research Paper

A Multivariate Test-Retest Approach for Verifying Self-Reported Speeding Behaviours among Taxi Drivers in Tehran

Hamid Reza Behnood[1,*], Sonja Forward[2], Jan Andersson[3]

Received: 2024/07/28 *Accepted: 2025/11/30*

Abstract

The aim of this study is to assess the repeated responses and verification of surveys using a one-way multivariate test-retest process with a four-week interval between measures for 120 taxi drivers in Tehran. The only independent variable includes the questioning stage. The items of a scenario-based questionnaire were considered as dependent variables. The null hypothesis posits no significant difference. The results were addressed using two testing approaches: variance testing and average testing. The results of the multivariate test statistics showed no significant variance difference in the answers between the first and the second test. The average testing including the contrast test also showed such a conclusion for the average measures in two stages. This means that factors which motivate speeding amongst taxi drivers were stable over a four-week period. In all these cases, the p-values were greater than 0.05, which indicates the non-significance of the difference between the responses in two stages, and as a result, accepting the null hypothesis. The added power in multivariate analysis results from the correlation between outcome variables and the use of cross-products.

Keywords: Reliability, MANOVA, Survey validation, Self-reporting, Theory of planned behaviour (TPB), Driver Behaviour Questionnaire (DBQ)

* Corresponding author. E-mail: hr.behnood@gmail.com
[1] Transportation Department, Faculty of Engineering, Imam Khomeini International University, Qazvin, Iran
[2] Swedish National Road and Transport Research Institute (VTI), Linköping, Sweden
[3] Swedish National Road and Transport Research Institute (VTI), Linköping, Sweden

1. Introduction

The use of driving behaviour models and theories have achieved significant progress in recent years, and fortunately, it has reached an acceptable level of practical application, especially in the field of risk assessment and road user education. A theory is a conceptual organization of concepts, mechanisms, and processes that are involved in the operation of a system, such as the atoms in a molecule or – of greater relevance here – driver in traffic [Shinar, 2017]. The Theory of Planned Behaviour (TPB) [Ajzen, 1991] and the Driver Behaviour Questionnaire (DBQ) [Reason et al., 1990] have been among the most common tools used to analyse the behaviour and the intentions of drivers leading up to crashes [Ajzen, 2020; Ziakopoulos et al., 2020]. These inventories are usually based on self-reporting data, so that on the one hand, questions are asked about their beliefs and norms and sometimes also if they have been involved in any crashes or near-crashes.

Self-reports and especially questionnaires have many advantages. They are usually less expensive than studies using an instrumented vehicle or a simulator, they provide more detailed information than observations, and they can reach large number of people [Lajunen & Özkan, 2011]. Lajunen and Özkan (2011) concluded that the gap between the reality and the responses given by self-reports may not be possible to erase completely, but at least it can be considerably reduced with adequate use of self-report methodology. Regardless of some of its drawbacks, studies show that researchers in general ‘trust’ self-reports [Kamaluddin et al., 2018].

However, even if self-reported data are usually trusted, it is important to assess its reliability but also the validation and verification to ensure the correctness of responses. That is, by re-asking the same questions to the same people after a certain time interval. In the past decades, various studies have addressed this approach in the form of test-retest techniques [Guttman, 1945; Faulstich, 1985; Waters et al., 1987].

Test-retest reliability is a measure of reliability obtained by administering the same test twice over a period of time to a group of individuals [Colin & Julie, 2005]. Meanwhile, in the field of driving behaviour studies, the validation of self-report questions related to mileage, violations and crashes has been of great interest [af Wåhlberg & Dorn, 2015; Shaikh et al., 2022; and Hill et al., 2023]. The methods used in test-retest validation processes mainly include univariate methods. But it is plausible that multivariate predictive models, which aggregates information, will produce larger effect sizes than univariate methods [Taxali et al., 2021]. In theory, multivariate analysis of variance (MANOVA) has greater power than analysis of variance (ANOVA) to detect effects because it takes account of the correlations between outcome variables [Hubert & Morris, 1989]. MANOVA has the power to account for correlation between outcome variables, and it does this by using cross-products [Field, 2017].

Self-reporting has been used as an inexpensive and useful tool in collecting information about driver behaviour in previous studies [Kamaluddin et al., 2018; Nguyen-Phuoc et al., 2021; Hwang and Choi, 2021; Hosseinpour et al., 2021; and Chouhan et al., 2022]. However, Af Wåhlberg and Dorn (2015) were less positive and argued that especially self-reported mileage, violations and crashes should be treated with extreme caution. Although multivariate analysis of variance (MANOVA) has been used in the analysis of human factors, driving behaviour and self-reported data [e.g. Davey et al., 2006; Taubman-Ben-Ari et al., 2016; Andersson et al., 2022], but test-retest processes using this type of analysis have been less used in such studies. However, this type of validation using multivariate analysis has been used in other sciences, especially in medical fields. For example, to show how latent growth models with planned incomplete data can be used to separate psychometric components of developmental interest, McArdle and Woodcock [1997] presented a brief introduction on components of change with test-retest data. Then, a time-lag design was introduced and illustrated with both short-term univariate and long-term multivariate studies of memory and reading. In a study about self-regulated learning in physical therapy education, van Lankveld et al. [2019] analysed the change in psychological capability and eagerness using a multivariate test-retest analysis of variance.

The aim of this study is to assess the repeated responses and verification of self-reports in a model loosely based on the TPB and the DBQ using a test-retest process, in which one-way multivariate analysis of the responses is used. This means that each question of the questionnaire is asked as a dependent variable at two different times. Thus, the only independent variable includes the questioning stage, which takes each dependent variable (each item of the questionnaire including behavioural items and crash records) in two stages. As a result, the only treatment of the study can be considered creating a time interval – which in this study is four weeks. The four-week interval was chosen as a base time and not adopted from a similar study. This duration was selected not too short to fade the response differences and not too long to possibly effect the drivers' behaviour. For this reason, the assumption of the correctness of answers can be expressed in such a way that the passage of time does not have a significant effect on the given answers or in such a way that there is no significant difference between the answers in the two stages of test and retest. In this regard, both average and variance measures will be examined. That means, it will be assessed whether there will be a significant change in the average of the answers with the passage of time or not, and in addition to this, what will be the change in the variance of the answers? This investigation will show whether the observed change in the variation of the responses will be the same as the change in the mean. For this purpose, we will use the process of multivariate analysis of variance (MANOVA) and the criteria for confirming the corresponding hypothesis test.

2. Methods

2.1. Participants

A preliminary set of 120 respondents who agreed to response the questions twice, including male taxi drivers in 52 taxi terminals in Tehran was used. The final sample of 120 participants in this study was obtained by revisiting the place of taxi terminals in which a previous set of 1000 data belonging to another study was collected. This way, 120 drivers who had previously responded to the questionnaire accepted to take part in the survey for the second time. In this study, the term 'taxi' refers to the cars under the supervision of Tehran Taxi Organization, hence ride-hailing and private taxis are outside the scope of the research. Descriptive data show that 72.8% of taxi drivers were over 42 years old and 43.1% were over 50 years old. Driving experiences for 65% of the drivers fall between 21 to 30 years and less than 7% with an experience below 10 years. After a four-week interval the same people responded to the same questionnaire again. The ethical issues such as informed consent and data confidentiality were considered in this study by not forcing the drivers to answer and keeping the anonymity of paper questionnaires.

2.2. Material

The questions were loosely based on the TPB and the DBQ. The first questions in the survey included questions measuring error and lapses, not violations. Violations were covered by questions based on the TPB. Errors were measured by 7 items (e.g., "Overtake someone that I had not noticed to be signalling a left turn" and "I misjudge my own speed when leaving the main road"). Lapses were measured by three items (e.g., "I get into the wrong lane when approaching the junction" and "When turning right at the cross-road, I use the wiper instead of the indicator”). In order to assess the participants attitudes, norms and perceived behaviour control they were presented with three different scenarios. The first one also included a photo, see Figure 1. In this scenario the principle of compatibility target context and time were clearly specified [see Ajzen & Fishbein, 2005].

Suppose you are driving your taxi on the left lane on Hafte-Tir Boulevard (pictured below) and you are approaching the Mesbah signalized cross-road. It is in the middle of the week and the time is around 11:00. The speed limit on this avenue is 40 km/h and in the distance, you can see that the traffic light is green. You want to cross the cross-road before it turns to red and therefore you increase your speed to 60 km/h. With this in mind, rate the following questions.

Figure 1. The scenario

The following questions were to be answered with this scenario in mind regardless of whether it would apply to them or not (1=strongly disagree, 5=strongly agree). The attitude towards the behaviour was measured by six items. Two of describing negative consequences (i.e., being frustrated; unnecessary risks) and for positive consequences (i.e., feel brave, accidents extremely low and being fined also extremely low). The combination of the indirect measures presented an alpha value of 0.823.

In the context of the above scenario, three item measuring normative beliefs were included that dealt with what their co-workers, other drivers and passengers thought about the behaviour. Descriptive norm was measured by one item mentioning what they believed other taxi drivers would do in the same situation. Control belief was measured using two items: easy to drive like this and more likely if they are in a hurry. The combination of the indirect measures presented an alpha value of 0.836. Intention to speed was measured by one item “I intend to drive in 60 km/h on a 40 km/h road within the next month”.

In scenario 2 the participants were asked to consider what would happen if they did not increase their speed to 60 km/h, as indicated in scenario 1, and stayed within the posted speed limit of 40 km/h. This was then followed by three items measuring subjective norm and if their passenger would think they were a slouchy driver, that the driver behind them and their co-worker would regard them as clumsy. The combination of the indirect measures presented an alpha value of 0.791.

In scenario 3 the participants were asked to imagine themselves in the same scenario as scenario 1 but this time they have a car in front of them who is driving in 40 km/h. This scenario was followed by four questions asking them if they would drive close to the car and blow the horn, if they would get very angry or very frustrated and if they would follow the other driver in order to retaliate.

All responses for scenarios 1 and 2 were measured on a Likert scale from 1 to 5 (I do not agree at all; I slightly agree; To some extent; I very much agree; I strongly agree) never, seldom, sometimes, often and very often). The third scenario was measured a five-point scale (Very unlikely; Unlikely; Somewhat likely; Likely; Very likely). The end of the survey included some background questions such as: age, education level, residence area type, driving experience, exposure and the number of crashes resulted in a fatal injury (last three years) or near crashes (last month).

Finally, the self-reported information related to driver's exposure and crash experiences includes the number of driving hours per day, the number of fatal, injury and property damage crashes, and near-crashes. After the final design of the questionnaire, its eloquence was controlled during interviews with five non-academic persons.

2.3. Procedure

This study includes initial steps of a larger study aimed at analysing driver behaviour loosely based on the TPB. In order to analyse the reliability of the questionnaire, a small sample of 90 taxi drivers were approached, of which only 53 people were willing to reply to the survey a second time. To increase the response rate a second group of drivers was contacted. Finally, the response from 120 drivers was included in the test-retest analysis.

To carry out the test-retest process, the questionnaire was given to the same 120 taxi drivers again after four weeks. All data related to drivers was collected in 52 taxi terminals located in 22 urban areas of Tehran. Many of these paper questionnaires were completed at CNG stations located near the terminals when the driver was stopped and idle for several minutes. Internet questioning was not used because the drivers would be ensured complete anonymity. Information of 120 participating drivers in the second stage (for the purpose of test-retest analysis) were collected through phone calls of some drivers who had entered their phone number (optionally) in the questionnaire or by returning to the terminals and allocating the same drivers again. To minimize the decline probability from drivers, a small gift including a pen was presented to the drivers.

2.4. Statistical Analysis

2.4.1. Data Reduction

The items in the questionnaire were subjected to principle-components factor analysis (PCA) and an exploratory factor analysis (EFA). To extract factors Kaiser normalization were used as the rotation method. By determining the factors, Cronbach's alpha was used as an internal consistency scale for each factor as well as the entire TPB questionnaire.

To reduce the number of independent variables in the analysis only the average of the group is included in the analysis. The groups are defined by the factor analysis in case of convincing result measures such as high Cronbach's alpha or KMO and Bartlett's test. In case of low satisfied performance measures, a nominal grouping of the items will be used. The average answer to questions related to errors and the average answer to questions related to lapses are entered separately. The value of the intention to violation is also the average answer of two questions in this field. Questions related to aggressive driving included four items, the average of which is included in the analysis. In the main part, the questions related to each factor extracted from PCA in the first-scenario TPB sub-inventory are placed in a group and the average answers are included in the analysis in case of high alpha values. Self-reported values for exposure criteria (number of daily driving hours), total number of crashes (sum of fatal, injury and PDO) and near-crash are also included in the analysis separately. By determining the reduced data by grouping them, the values taken in two stages with a time interval of four weeks for each respondent are then placed next to each other.

2.4.2. Multivariate Test-Retest

A two-step questioning process was used to analyse test-retest reliability. For this purpose, from a part of the same sample volume on which exploratory factor analysis was performed, after a time interval of four weeks, the questions of the same questionnaire were asked again. This analysis includes a hypothesis test in which the independent variable includes the questioning stage and the dependent variables include the responses of each item or their group average. The null hypothesis is expressed in such a way that there is no significant difference between the responses in two stages. In other words, the respondents gave the same answer in the second stage as they did in the first stage.

The criteria used in this analysis follow two approaches. One is that during the two stages of questioning, what changes have been made in the average responses of each group or item, and the other is whether the variance of the data has changed in the test stage when compared to the retest stage or not. Examining the change in variance is done using criteria such as Box's Test, Levene's test, and quadruple tests of multivariate analysis including Wilk's Lambda, Pillai's trace, Hotelling's trace, and Roy's largest root. On the other hand, contrast analysis examines the difference in averages.

Tests of the equality of responses with multiple dependent variables are performed accordingly. First, the Box's Test statistic is controlled. Box's Test [Box, 1971] for equivalence of covariance matrices is a parametric test used to compare variation in multivariate samples and tests the null hypothesis that the observed covariance matrices of the dependent variables are equal across groups. Tests that are best for detecting small deviations from the null hypothesis are derived for a number of hypotheses concerning multivariate normal populations [John, 1971; and Davidson, 1972]. Multivariate testing is a technique for testing a hypothesis in which multiple variables are modified. The four statistics used in MANOVA include Wilk's Lambda, Pillai's trace, Hotelling's trace, and Roy's largest root [Field, 2017]. Levene's test is the usual *F*-test for equality of means computed on pseudo-observations, which one defines as the absolute deviations of the data points from an estimate of the group 'center' [Carroll & Schneider, 1985]. It tests the null hypothesis that the error variance of the dependent variable is equal across groups. The residual (or error) sum of squares and cross-products (SSCP) matrix shows the degree of correlation among the dependent variables. It represents the ratio of systematic to unsystematic variance in the model [Field, 2017].

Contrast analysis is a relatively simple but effective statistical method for testing theoretical predictions about differences between group means against the empirical data [Haans, 2018]. The univariate test confirms the multivariate test results, in the sense that there is no significant difference between independent variables to affect the dependent variables.

3. Results

The results of the research include two parts: data reduction and the multivariate analysis of the answers in the two stages of test and retest. In the test-retest analysis, the results of the multivariate approach have been compared with common univariate methods, including correlation and error measures between the two response stages.

3.1. Data Reduction

The amount of factor loadings related to each of the items assigned to each factor were calculated for the first TPB scenario and the DBQ (i.e., errors and lapses). Cronbach's alpha values were not within acceptable limits for some explored factors. The TPB factor analysis resulted four factors with alpha values of 0.746, 0.938, 0.682, and 0.046 with the total value of 0763. The DBQ factor analysis included three factors with alpha values 0.693, 0.612, and 0.306 and a total value of 0.719. In addition, the KMO and Bartlett's Test measure of sampling adequacy (0.668 for TPB scenario 1 and 0.719 for errors and lapses) for 120 respondents did not support a high level of acceptance for the two groups.

Due to the poor results obtained in PCA analysis, it cannot be concluded that data reduction can be done based on exploratory factors. Therefore, to categorize the data, it is inevitable to assign the items to a nominal group based on their meaning. For this purpose, both categories of questions related to TPB scenario 1, 2 and 3 and errors and lapses are included in Tables 1 and 2 respectively according to the nature of the question. To this end, 12 items were identified in three scenario groups. Scenario 1 includes attitude, perceived control and norms for speedy driving. Scenario 2 includes subjective norms for driving according to the posted speed limit of 40 km/h and failing to pass the green light before it turns red. Finally, Scenario 3 includes four items related to aggressive driving. Group names for errors and lapses in Table 2 reflect the nature of errors, i.e., deficiency in knowledge of traffic rules, and lapses, i.e., forgetting or lack of concentration about them.

Table 1. Naming the groups of scenario-based items

Group name	Items: scenario 1
Affective attitude	Q2.1. If I can cross the cross-road before it turns to red, I feel brave. Q2.5. If I failed to cross the cross-road before the light changed to red, I would feel very frustrated.
Risk	Q2.2. If I increase my speed to 60 km/h then I take unnecessary risks.
Positive attitude towards speeding	Q2.3. To be involved in an accident with another vehicle is extremely low. Q2.4. To be involved in an accident with a pedestrian is extremely low. Q2.6. If I drive in 60 km/h then the risk of being fined by the police is extremely low.
Subjective norm	Q2.7. If I drive at a speed of 60 km/h the passenger thinks that it is acceptable. Q2.8. If I drive at a speed of 60 km/h other taxi drivers thinks that it is acceptable.
Passenger and affective	Q2.9. If I drive at a speed of 60 km/h the passenger would get worried.
Perceived control	Q2.10. For me it would be very easy to drive in 60 km/h. Q2.11. I would be more likely to increase the speed to 60 km/h if I was in a hurry.
Descriptive norm	Q2.12. Other taxi drivers would increase their speed to 60 km/h if they wanted to cross the cross-road before the light turns to red.
	Items: scenario 2
Subjective norm	Q3-1. If I do not increase the speed to 60 km/h and fail to pass the green light then the passenger thinks that I am a slouchy driver. Q3-2. If I do not increase the speed to 60 km/h and fail to pass the green light, then the driver behind me considers me a clumsy driver. Q3-3. If my co-workers see that I fail to pass the green light, they think I am a clumsy driver.
	Items: scenario 3
Aggressive driving	Q4-1. Under the circumstances, I get very close to the car in front and blow the horn to make her/him move faster or get out of my way so I can get through the green light quickly. Q4-2. If a car stops me from increasing my speed, I will get very angry Q4-3. If a car stops me from increasing my speed, I will feel very frustrated Q4-4. If a driver has not moved out of the way even if I wanted him/her to do so, I will follow her/him after the traffic light to retaliate.

Table 2. Naming the group of items in errors and lapses

Group name	Items
Errors	Q1.1. Overtake someone that I had not noticed to be signaling a left turn.
	Q1.2. When queuing to turn left onto a main road, I fail to pay close attention to the mainstream traffic and as a result I nearly hit the car in front.
	Q1.3. When turning left at the cross-road, I fail to notice two wheelers.
	Q1.4. I forget to look in the rear mirrors when I change lanes.
	Q1.8. I misjudge my own speed when leaving the main road.
	Q1.9. I misjudge the speed of other drivers on the main road when wanting to enter the same road.
	Q1.10. I fail to notice that the traffic light has turned red.
Lapses	Q.1.5. I get into the wrong lane when approaching the junction.
	Q.1.6. I misunderstand the signs and therefore end up in the wrong direction at the cross-road.
	Q.1.7. When turning right at the cross-road, I use the wiper instead of the indicator.

3.2. Simple Test-Retest

By grouping the items, it became possible to reduce the data to 15 variates. Table 3 shows the classification of the mentioned data as dependent variables and the mean and standard deviation of their answers in two stages of test and retest. In Table 3 results from questions about crashes, near crashes and exposure are also included. In most cases, a slight increase in the average value of responses can be seen. Among the scenario 1 TPB factors, passengers and affective factor has the highest average value in both stages. As expected, and following the nature of randomness and the rarity of crashes, the value of the standard deviation and following that the variance of the crashes have exceeded its average. The same case noticeably applies for the near-crash data. Test-retest correlation values show a conventional measure to show the correlation of answers in two stages. Here, due to the use of average factors and data groups and the continuity of these average values, Pearson correlation has been used. Based on this simple criterion, all the answers have a high correlation. Now we control this result with the root-mean-square error (RMSE) criterion. This criterion is a frequently used measure of the differences between values (sample or population values) predicted by a model or an estimator and the values observed [Hyndman & Koehler, 2006]. The lower values of this criterion indicate the closer the responses of the two stages, but considering that the data was measured on different scales, it is not possible to directly compare the RMSE values, and preferably we use its normalized value, that is, the coefficient of variation of RMSE (CV(RMSE)=RMSE/Test-retest mean value). In Table 3, both RMSE and CV(RMSE) values for the near-crashes variable are higher than other variables. This is also confirmed by the mean absolute error (MAE) measure. Such a result by RMSE and MAE measures shows that the responses in the near-crash data variable had the greatest difference in the two response stages (CV(RMSE)=1.14 and MAE=0.89), while the test-retest correlation criterion (=0.91) showed the opposite result (a lower difference).

Table 3. Descriptive statistics for test-retest data

Dependent variable	Stage 1 (test)		Stage 2 (retest)		Test-retest correlation	Test-retest RMSE	Test-retest CV(RMSE)	Test-retest MAE
	Mean	SD	Mean	SD				
Affective attitude	2.05	0.93	2.14	0.92	0.90	0.41	0.20	0.24
Risk	3.20	1.26	2.93	1.20	0.90	0.62	0.20	0.35

Dependent variable	Stage 1 (test)		Stage 2 (retest)		Test-retest correlation	Test-retest RMSE	Test-retest CV(RMSE)	Test-retest MAE
	Mean	SD	Mean	SD				
Positive attitude towards speeding	2.71	1.06	2.66	1.02	0.96	0.31	0.12	0.19
Subjective norm	2.89	1.01	2.92	1.02	0.95	0.34	0.12	0.15
Passenger and affective	3.53	1.18	3.56	1.18	0.95	0.47	0.10	0.13
Perceived control	3.07	0.94	3.06	0.95	0.89	0.44	0.14	0.26
Descriptive norm	2.80	1.05	2.88	1.01	0.85	0.56	0.20	0.28
Subjective norm (Scenario 2)	2.19	0.95	2.24	0.97	0.86	0.50	0.23	0.23
Intention	2.43	1.09	2.45	1.10	0.92	0.44	0.18	0.19
Errors	1.86	0.55	1.93	0.54	0.85	0.31	0.16	0.20
Lapses	1.47	0.43	1.56	0.56	0.80	0.35	0.23	0.19
Aggressive driving	1.86	0.77	1.94	0.79	0.92	0.27	0.14	0.17
Exposure (hr/day)	9.55	2.27	9.72	2.07	0.91	0.96	0.10	0.38
Total crashes	0.74	1.02	0.83	1.01	0.93	0.39	0.49	0.15
Near-crashes	2.54	5.71	2.13	4.03	0.91	2.67	1.14	0.89

3.3. Multivariate Variance Testing

In the following, the results of the test-retest study with a multivariate approach on 120 repeated responses are presented. It is reminded that the null hypothesis is tested in this way: There are no differences between test (stage 1) and retest (stage 2) responses. The independent variables are the two stages 1 (test) and 2 (retest) and the dependent variables include 15 variates which were introduced in Table 3 and are simultaneously included in the analysis. In the first step, Box's Test of Equality of Covariance Matrices is performed. The p-value for the F statistic in this test is equal to 1.00. Thus, the significance level for F is well above 0.05, so the null hypothesis is accepted and we observe the equality of covariance matrices for dependent variables across different groups of dependent variables.

Table 4 shows the results of the multivariate test for the independent variable 'stage'. The 'stage' effect shows the effect or lack of effect of the passing of time on the responses. The results of four multivariate test statistics and their significance values are presented in this table, as described in section 2.4.2. In all these cases, the p-value criterion equalling .954 (>0.05) obtained, which indicates the non-significance of the difference between the responses in two stages, and as a result, accepting the null hypothesis. In other words, no significant change has been observed in the answers given by the drivers over time. Among the statistics of Pillai's Trace and Wilks's Lambda, the latter value is closer to one, which shows the significant similarity of the responses in the two stages. Partial eta squared is telling us how large of an effect the independent variable(s) had on the dependent variable. The low values for this criterion (.031) show small effect of the time interval on the response values.

Table 4. Multivariate test results for the Stage effect

Criterion	Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
Pillai's Trace	.031	.470	15.000	224.000	.954	.031
Wilks' Lambda	.969	.470	15.000	224.000	.954	.031
Hotelling's Trace	.031	.470	15.000	224.000	.954	.031

Criterion	Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
Roy's Largest Root	.031	.470	15.000	224.000	.954	.031

The Levene's test showed that the significance levels of *F* statistic for all dependent variables are above 0.05 except for the 'lapses' (Sig.=0.001), so the error variances for the two stages are equal except for the 'lapses' group. It strengthens the case for assuming that the multivariate test statistics are robust. Of course, the higher values of the F statistic for the average lapses, TPB 'Risk' item and near-crash variables indicate a greater error variance between two stages than the variance within each stage. This issue shows the difference in most of the answers for these variables, but in any case, this difference is not significant. Such a result has also been observed in tests of between-subjects effects. This test shows that the Stage effect on all dependent variables is not significant. As well, the adjusted R squared values (all near zero) show a small amount of the Stage effect on dependent variables.

Finally, the near-zero values of hypothesis-error ratio or in other words the ratio of systematic to unsystematic variance in the model for the SSCP matrix show a very small effect of variables in the model.

3.4. Multivariate Average Testing

Univariate analysis by considering the variables separately can reveal the magnitude of the error, i.e., the difference between the two stages. This analysis somehow evokes the same dependent variables used in Table 3. The contrast analysis in Table 5 in line with the results of univariate analysis, shows the difference between the averages of the two groups in more detail. The three estimation criteria of contrast, difference and standard error simultaneously shows the close order of the answers in two stages. Here, more robustly, we can conclude that near-crash, exposure, and the TPB 'Risk' variables had the greatest difference in two stages, however according to the obtained p-values, we can see these differences are still not significant.

4. Discussion

The results were addressed using two testing approaches: variance testing and average testing. The results of the multivariate test statistics, including Box's test and Pillai's Trace, Wilks' Lambda, Hotelling's Trace and Roy's Largest Root showed no significant variance difference in the answers between the first and the second test. The average testing including the contrast test also showed such a conclusion for the average measures in two stages. This means that factors which motivate speeding amongst taxi drivers was stable over a four-week period. The Levene's test showed that the error variances for the two stages were equal except for the factor measuring lapses, which strengthens the case that the multivariate test statistics are robust. In all these cases, the p-value criterion greater than 0.05 was obtained, which indicates the non-significance of the difference between the responses in two stages, and as a result, accepting the null hypothesis. In other words, no significant change has been observed in the answers given by the drivers over time. The added power in multivariate analysis results from the correlation between outcome variables and the use of cross-products. Regarding the results related to the TPB scenario 1 groups, in both univariate and multivariate analysis, the difference in responses in the two stages was very small. It should be mentioned that some studies that have used the simple correlation approach have shown more but still insignificant differences in the answers related to these factors [e.g., Piazza et al., 2019; and Qiao et al., 2021].

The results from this study show a simple but logical and robust sample of two questioning stages with a time interval of four weeks. In the simple analysis using conventional criteria such as test-retest correlation and RMSE, both RMSE and CV(RMSE) values for the near-crashes variable were higher than other variables. This was also confirmed by the mean absolute error (MAE) measure. Such a result by

RMSE and MAE measures showed that the responses in the near-crash data variable had the greatest difference in the two response stages, while the test-retest correlation criterion showed the opposite result (a lower difference). This conclusion shows that the application of simple criteria in univariate analysis might be inconclusive and in some cases they are misleading. That is, we saw a contradiction in the interpretation of the results, which was resolved by performing multivariate analysis. This added power in multivariate analysis, as pointed out by various researchers [Hubert & Morris, 1989; Field, 2017; and Taxali et al., 2021], results from the correlation between outcome variables and the use of cross-products. The multivariate test-retest reliability analysis was conducted for variance testing and average testing, both of which showed that the differences between two stages was not significant. In this analysis, it was observed that despite the insignificant difference between the test and retest stages, some variables such as near-crashes and exposure showed a higher level of difference than other variables. af Wåhlberg and Dorn (2015) pointed out that exposure (i.e., mileage) information can be subject to a level of subjectivity, unlike crash record information, and therefore its reliability is affected by time. The reason for the results could be explained as follows:

Table 5. Contrast analysis results

Stage Simple Contrast[a]		Affective attitude	Risk	Positive attitude towards speeding	Subjective norm	Passenger and affective	Perceived control	Descriptive norm	Subjective norm (scenario 2)	Intention	Errors	Lapses	Aggressive driving	Exposure (hr/day)	Total Crashes	Number of near crashes
		Dependent Variable														
Contrast Estimate		-.092	.267	.042	-.033	-.033	.004	-.083	-.050	-.025	-.069	-.089	-.079	-.167	-.083	.408
Hypothesized Value		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Difference (Estimate - Hypothesized)		-.092	.267	.042	-.033	-.033	.004	-.083	-.050	-.025	-.069	-.089	-.079	-.167	-.083	.408
Std. Error		.120	.159	.135	.132	.153	.122	.133	.124	.142	.071	.064	.101	.281	.131	.640
Sig.		.445	.096	.759	.800	.828	.973	.532	.688	.861	.328	.169	.433	.554	.526	.524
95% Confidence Interval for Difference	Lower Bound	-.328	-.047	-.225	-.293	-.334	-.236	-.346	-.295	-.305	-.208	-.216	-.278	-.721	-.342	-.853
	Upper Bound	.144	.581	.308	.226	.268	.245	.179	.195	.255	.070	.038	.119	.388	.175	1.670

First, it might be difficult for the drivers to remember exactly how much they have been driving and the same with near-crashes, especially for a taxi driver. Second, the first time they were asked about this could have made them to start thinking and then concluding that they perhaps were wrong. Another explanation could be that after four weeks they had been driving more hours per day. As a different result, unlike some studies in which using simple test-retest, a small correlation was obtained for the variable of intention to violation [e.g., Sakashita et al., 2014], the present study using multivariate test-retest found a low level of difference for this variable. Though, some studies using simple test-retest have shown a lower level of difference between two stages for such a variable [e.g., Guo et al., 2013]. Although, given the limits of their study, it is difficult to draw any firm conclusions about how strong systematic response bias and random error effects are in crash data, not to mention self-reported violations and mileage. Several researchers have tried to address the reliability for various variables through different approaches. Kamaluddin et al. (2018) mapped the current practice (including 134 studies worldwide) in the collection of road traffic crash data by surveying studies where traffic crashes were reported by the involved road users. In some studies [e.g., Begg et al., 1999; and Arthur et al., 2005], the comparison of self-reporting results with real data recorded in crash or traffic violation databases has been used. Of course, many data, such as behavioural and cognitive characteristics, cannot be observed and validated using databases, such as information including items in common and well-known inventories that can only be obtained by questioning [e.g., the studies conducted by Carlson et al., 2011, Shen et al., 2018, Bernstein & Calamia, 2019 and Chouhan et al., 2022]. Accordingly, to produce knowledge-based safety measures it could be argued that developing countries should increase their efforts when it comes to using self-reporting to understand what predicts driver behaviour.

5. Conclusion

In this study, self-reported information from 120 taxi drivers in Tehran was verified through repeated questioning including items loosely based on TPB and DBQ, using a test-retest process. For this purpose, each question in the questionnaire was asked at two different occasions with a four-week interval between measures. The content of the scenario-based questionnaire for this study consists of questions related to errors and lapses (DBQ), items included in the TPB, aggressive driving, amount of daily driving and crash and near-crash records. Due to the weak results obtained in PCA analysis, the data reduction could not be done based on exploratory factors. Therefore, in order to categorize the data, it was inevitable to assign the items to a nominal group based on their meaning. The results of four multivariate test statistics, including Pillai's Trace, Wilks' Lambda, Hotelling's Trace and Roy's Largest Root, as well as their significance values, were presented and the results showed no significant difference in the answers in two stages which meant that the self-reported survey was reliable. The Levene's test showed that the error variances for the two stages were equal except for the factor measuring lapses. Hence, we more strongly assume that the multivariate test statistics are robust. For the average testing approach, the contrast analysis showed the difference between the averages of the two groups in more detail. As a supporting point univariate analysis, which consider the variables separately, can reveal the magnitude of the error.

In this analysis, it was observed that despite the insignificant difference between the test and retest stages, some variables such as near-crashes and exposure showed a higher level of difference than other variables. This study presented a new approach that can show a more robust image of drivers' answers in order to search for the real relationship between several stages of questioning in studies based on self-reporting. Therefore, it is suggested that future studies use such an approach, especially in models whose subjects have a lot of frequency and diversity of behaviour. This study showed that the main concern about the verification of data related to near-crash and exposure is more than other data there is a risk that sometimes the difference observed in them becomes significant. For this reason, it is recommended that in the studies, a reliable experiment design for this type of information must be foreseen.

6. Limitations

Although the criteria used in this analysis showed enough sample size, a larger database could have developed yet more reliable results. The reason for the smaller sample size in such studies is the difficulty to motivate respondents to take part in two similar studies, i.e., the retest stage. Another limitation in self-reported surveys is the possibility for the lack of trust from respondents. For this reason, the surveyors tried to improve confidence by keeping the anonymity of paper questionnaires. The final issue that limits the use of this research is its lack of generalizability. For another independent sample group, the analysis procedure should be repeated indeed.

7. References

– Af Wåhlberg, A.E. and Dorn, L., 2015. How reliable are self-report measures of mileage, violations and crashes? Safety Science, 76, pp.67-73.

– Ajzen, I., 1991. The theory of planned behavior. Organizational behavior and human decision processes, 50(2), pp.179-211.

– Ajzen, I., 2020. The theory of planned behavior: Frequently asked questions. Human Behavior and Emerging Technologies, 2(4), pp.314-324.

– Andersson, J., Bro, T. and Lajunen, T., 2022. Self-perception and the relation to actual driving abilities for individuals with visual field loss. Frontiers in human neuroscience, 16.

– Arthur Jr, W., Bell, S.T., Edwards, B.D., Day, E.A., Tubre, T.C. and Tubre, A.H., 2005. Convergence of self-report and archival crash involvement data: A two-year longitudinal follow-up. Human Factors, 47(2), pp.303-313.

– Begg, D.J., Langley, J.D. and Williams, S.M., 1999. Validity of self-reported crashes and injuries in a longitudinal study of young adults. Injury Prevention, 5(2), pp.142-144.

– Bernstein, J.P. and Calamia, M., 2019. Dimensions of driving-related emotions and behaviors: An exploratory factor analysis of common self-report measures. Accident Analysis & Prevention, 124, pp.85-91.

– Box, M.J., 1971. Bias in nonlinear estimation. Journal of the Royal Statistical Society: Series B (Methodological), 33(2), pp.171-190.

– Carlson, E.B., Smith, S.R., Palmieri, P.A., Dalenberg, C., Ruzek, J.I., Kimerling, R., Burling, T.A. and Spain, D.A., 2011. Development and validation of a brief self-report measure of trauma exposure: the Trauma History Screen. Psychological assessment, 23(2), p.463.

– Carroll, R.J. and Schneider, H., 1985. A note on Levene's tests for equality of variances. Statistics & probability letters, 3(4), pp.191-194.

– Chouhan, S.S., Kathuria, A. and Sekhar, C.R., 2022. Evaluating the correlation between risky riding behaviour and self-reported crashes in India: minimizing unobserved heterogeneity. IATSS research, 46(4), pp.515-524.

– Colin, P. and Julie, W., 2005. Exploring reliability in academic assessment. Graduate Assistants, UNI Office of Academic Assessment, College of Humanities and Fine Arts Student Outcomes Assessment.

– Davey, J., Freeman, J. and Wishart, D., 2006. A study predicting self-reported crashes among fleet drivers. In 2006 Australasian Road Safety Research, Policing and Education Conference Proceedings (pp. 1-11). Able Video and Multimedia Pty Ltd.

– Davidson, M.L., 1972. Univariate versus multivariate tests in repeated-measures experiments. Psychological Bulletin, 77(6), p.446.

– Faulstich, Michael E., 1985. The reliability of psychophysiological assessment: a comparative analysis of two mental and two physical stressors (test-retest reliability, physiological reactivity, cold pressor, blood pressure, heart rate)." PhD diss., Louisiana State University and Agricultural & Mechanical College.

– Field, A., 2017. Discovering statistics using IBM SPSS statistics. sage.

– Guo, F., Simons-Morton, B.G., Klauer, S.E., Ouimet, M.C., Dingus, T.A. and Lee, S.E., 2013. Variability in crash and near-crash risk among novice teenage drivers: a naturalistic study. The Journal of pediatrics, 163(6), pp.1670-1676.

– Guttman, L., 1945. A basis for analyzing test-retest reliability. Psychometrika, 10(4), pp.255-282.

– Haans, A., 2018. Contrast analysis: A tutorial. Practical Assessment, Research, and Evaluation, 23(1), p.9.

– Hubert, C. and Morris, J., 1989. Multivariate versus multiple univariate analysis. Psychological Bulletin, 105, pp.302-308.

– Hyndman, R.J. and Koehler, A.B., 2006. Another look at measures of forecast accuracy. International journal of forecasting, 22(4), pp.679-688.

– John, S., 1971. Some optimal multivariate tests. Biometrika, 58(1), pp.123-127.

– Kamaluddin, N.A., Andersen, C.S., Larsen, M.K., Meltofte, K.R. and Várhelyi, A., 2018. Self-reporting traffic crashes–a systematic literature review. European transport research review, 10(2), pp.1-18.

– Lajunen, T. and Özkan, T., 2011. Self-report instruments and methods. In Handbook of traffic psychology (pp. 43-59). Academic Press.

– McArdle, J.J. and Woodcock, R.W., 1997. Expanding test–retest designs to include developmental time-lag components. Psychological Methods, 2(4), p.403.

– Piazza, A.J., Knowlden, A.P., Hibberd, E., Leeper, J., Paschal, A.M. and Usdan, S., 2019. Mobile device use while crossing the street: Utilizing the theory of planned behavior. Accident Analysis & Prevention, 127, pp.9-18.

– Qiao, X., Ji, L., Jin, Y., Si, H., Bian, Y., Wang, W. and Wang, C., 2021. Development and validation of an instrument to measure beliefs in physical activity among (pre) frail older adults: an integration of the health belief model and the theory of planned behavior. Patient Education and Counseling, 104(10), pp.2544-2551.

– Reason, J., Manstead, A., Stradling, S., Baxter, J. and Campbell, K., 1990. Errors and violations on the roads: a real distinction?. Ergonomics, 33(10-11), pp.1315-1332.

– Sakashita, C., Jan, S., Senserrick, T., Lo, S. and Ivers, R., 2014. Perceived value of a motorcycle training program: the influence of crash history and experience of the training. Traffic injury prevention, 15(4), pp.407-413.

– Shen, B., Qu, W., Ge, Y., Sun, X. and Zhang, K., 2018. The relationship between personalities and self-report positive driving behavior in a Chinese sample. PLoS one, 13(1), p.e0190746.

– Shinar, D. ed., 2017. Traffic safety and human behavior. Emerald Group Publishing.

– Taxali, A., Angstadt, M., Rutherford, S. and Sripada, C., 2021. Boost in test–retest reliability in resting state fMRI with predictive modeling. Cerebral cortex, 31(6), pp.2822-2833.

– Taubman–Ben-Ari, O., Eherenfreund–Hager, A. and Prato, C.G., 2016. The value of self-report measures as indicators of driving behaviors among young drivers. Transportation research part F: traffic psychology and behaviour, 39, pp.33-42.

– Van Lankveld, W., Maas, M., van Wijchen, J., Visser, V. and Staal, J.B., 2019. Self-regulated learning in physical therapy education: a non-randomized experimental study comparing self-directed and instruction-based learning. BMC medical education, 19(1), pp.1-9.

– Waters, W.F., Williamson, D.A., Bernard, B.A., Blouin, D.C. and Faulstich, M.E., 1987. Test-retest reliability of psychophysiological assessment. Behaviour research and therapy, 25(3), pp.213-221.

– Ziakopoulos, A., Tselentis, D., Kontaxi, A. and Yannis, G., 2020. A critical overview of driver recording tools. Journal of safety research, 72, pp.203-212.