Exploitation of naturalistic driving data to estimate crash risk through machine learning techniques

Spyros Tsigos[1], Eva Michelaraki[1*], Elena Provatari[1], George Yannis[1]

[1]National Technical University of Athens, Department of Transportation Planning and Engineering, 5 Heroon Polytechniou str., Athens, GR-15773, Greece
*Corresponding author: evamich@mail.ntua.gr

Abstract. Despite continuous investment in road and vehicle safety, as well as improvements in technology standards, the total amount of road crashes has been increasing over the last decades. The aim of the current study was to identify the most critical precursors of risk from both the task complexity and the coping capacity side. To that end, data collected from 30 drivers who participated in a naturalistic driving experiment for four months in UK were collected and analyzed. Several Structural Equation Models (SEMs) were applied in order to explore the effect between the latent variables of task complexity, coping capacity and risk. Results indicated that demographic characteristics, such as gender and age had a negative correlation with coping capacity indicating that male drivers and especially elderly people had a lower level of coping capacity, Moreover, better general driving skills were associated with higher coping capacity. It was also revealed that vehicle strain (increased vehicle age) along with type of fuel and trip difficulty were associated with higher task complexity levels. Overall, it is important to consider the specific factors and context involved when assessing the relationship between task complexity, coping capacity and risk.

Keywords: Task Complexity; Coping Capacity; Naturalistic Driving Experiment; Structural Equation Models; Road Safety.

1 Introduction

In modern society and reality, road transport is an integral part of people's daily lives, as they use it for all their activities during the course of a day. With the global increase in road transport, road crashes are also increasing, bringing to the surface the critical issue of road safety. According to the World Health Organization, annual road traffic fatalities amount to 1.35 million and are the leading cause of death in people aged 5-29 years, with 1 in 2 victims being pedestrians, motorcyclists or cyclists (WHO, 2018).

The aim of the current study was to identify the most critical precursors of risk from both the task complexity and the coping capacity side. To that end, data collected from 30 drivers who participated in a naturalistic driving experiment for four months in UK were collected and analyzed.

The paper is set up as follows. The project and its overarching goal are thoroughly described at the outset. After that, a thorough background on the statistical methods used to analyze driving behavior is offered. The research methodology is then explained, along with the models' theoretical underpinnings. Finally, the results are discussed, followed by important conclusions about how crucial elements like task complexity and coping capacity affect risk.

2 Background

To begin with, task complexity relates to the current status of the real world context in which a vehicle is being operated. In particular, task complexity context is monitored via registration of road layout (i.e. highway, rural, urban), time and location, traffic volumes (i.e. high, medium, low) and weather. On the other hand, coping capacity is dependent upon two underlying factors and it consists of several aspects of both vehicle and operator state. More specifically, the latent variables associated to "vehicle state" are estimated on the basis of various metrics, such as technical specifications, actuators & admitted actions and current vehicle status. Additionally, the latent variables associated to "operator state" are estimated on the basis of several aspects, such as mental state, driving behavior, competencies, personality, sociodemographic profile and health status. Figure 1 illustrates the conceptual framework for the prediction of risk in function of coping capacity and task complexity.

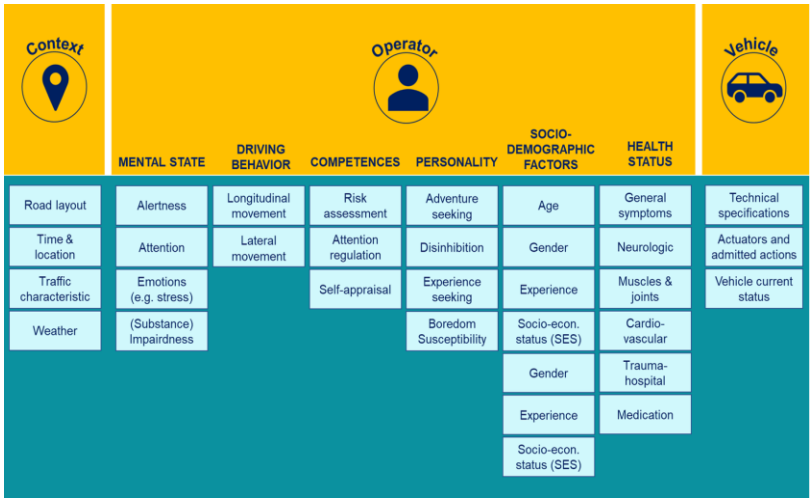

Figure 1: Prediction of risk in function of coping capacity and task complexity

The research by Ofori et al. (2023) examined the safety procedures and performance of Ghana's crucial oil, gas, and related energy sector. The aim of the work was to present an integrated approach for examining new antecedent metrics for improving the safety performance using partial least squares structural equation modelling (PLS-SEM). The outcomes showed that efficient safety training is another safety management technique that can aid in creating a successful safety policy. Nevertheless, neither safety training nor safety compliance had any discernible impact on safety performance.

The study by Useche et al. (2021), examined the risky behaviour of pedestrians on a road network by separating them into males and females using Structural Equation Models (SEM). This study investigated the influence of gender, both on violations of the Road Traffic Code, and on dangerous behaviors committed unintentionally, by applying SEM. According to the researchers, in male pedestrians, latent risk perception and educational level played a significant role in committing violations.

3 Methodology

3.1 Structural Equation Models (SEMs)

SEM is frequently used to describe intricate and multi-layered relationships between variables that are observed and those that are not, like "task complexity," etc. Unobserved variables are latent constructs, which are comparable to the components in a factor / principal component analysis, whereas observed variables are quantifiable.

A measurement model and a structural model make up structural equation models. The measurement model is used to calculate the associated measurement errors as well as how well different observable exogenous variables can measure (i.e. stress on) the latent variables. The SEM is used to investigate the relationships among the model variables, enabling the modeling of both direct and indirect linkages. SEMs are distinct from ordinary regression approaches where relationships among variables are direct.

The general formulation of SEM is as follows (Washington et al., 2020):

$$\eta = \beta\eta + \gamma\xi + \varepsilon \quad (1)$$

where η is a vector of endogenous variables, ξ is a vector of exogenous variables, β and γ are vectors of coefficients to be estimated, and ε is a vector of regression errors. The measurement models are then as follows (Chen, 2007):

$$x = \Lambda x\xi + \delta, \text{ for the exogenous variables} \quad (2)$$

$$y = \Lambda y\eta + \zeta, \text{ for the endogenous variables} \quad (3)$$

where x and δ are vectors which denote the observed exogenous variables and their errors, y and ζ are vectors which denote the observed endogenous variables and their errors, and Λx, Λy are structural coefficient matrices for the effects of the latent exogenous and endogenous variables on the observed variables. A path analysis, which illustrates how a group of 'explanatory' variables might have an impact on a 'dependent' variable, is frequently used to portray the structural model.

In the context of model selection, model Goodness-of-Fit measures consist an important part of any statistical model assessment. Several metrics are commonly used for the model evaluation. More specifically, the aforementioned metrics are the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), the Goodness-of-Fit Index (GFI), the (standardized) Root Mean Square Error Approximation (RMSEA), the Comparative Fit Index (CFI), the Tucker-Lewis Index (TLI) and the Hoelter index.

The latent variables and their constructs, used in this paper are the product of the authors own work and have been selected after continuous testing based on the optimum efficiency. A critical step in the operation of SEM models is the clustering of variables, in order to provide useful information and conclusions about latent variables. Following the principles of sequencing and hermeneutics, the clustering of variables was carried out. In particular, four latent variables were formed: i.e. task complexity, coping capacity-operator state, coping capacity-vehicle state and synthesis of risk factors.

The latent variables represent unobserved variables. A regression is then performed between the unobserved and observed independent variables, which are finally correlated with several dependent variables. Through tests carried out, the dependent variables for which the most significant results were obtained were speeding and harsh braking. The

results of the models were evaluated by satisfying the following statistical tests: p-value<0.001, CFI > 0.90, TLI > 0.90, SRMR<0.05 and RMSEA<0.05.

4 Results

4.1 Speeding

Risk is measured by means of the STZ (Safety Tolerance Zone) levels for speeding (level 1 refers to 'normal driving' used as the reference case, level 2 refers to 'dangerous driving' while level 3 refers to 'avoidable accident driving'), with positive correlations of Risk with the STZ indicators. Figure 2 summarizes the model fit of SEM applied for speeding. The numbers on the arrows indicate the correlation between the variables.

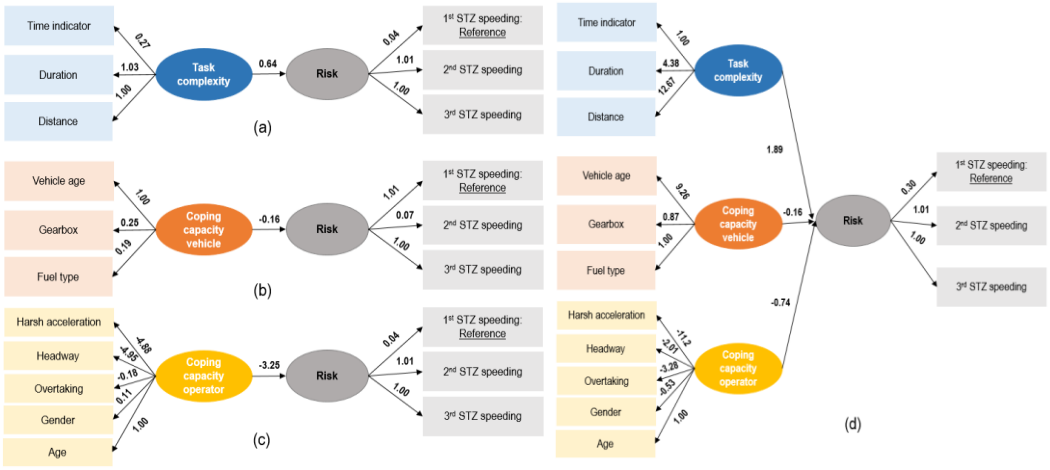

Figure 2. Results of SEM for speeding; experiment phase 1 (a), 2 (b), 3 (c), 4 (d)

The latent variable task complexity is measured by means of the environmental indicators "Time indicator". The exposure indicator of trip duration was also included in the task complexity analysis. The most important finding is the positive correlation between the complexity of the driving task and risk, meaning that increased task complexity relates to increased risk according to the model (regression coefficient=0.64).

The latent coping capacity is measured by means of operator state indicators, such as harsh acceleration, headway, illegal overtaking, age and gender. At the same time, the indicators of coping capacity - vehicle state, such as vehicle age, gearbox or fuel type are included in the SEM applied. It is noted that the structural model between coping capacity and risk shows a negative coefficient, which means that increased coping capacity relates to decreased risk according to the model (regression coefficient= -0.16 (vehicle state) / -3.25 (operator state).

Overall, the structural model between task complexity and risk shows a positive coefficient, which means that increased task complexity relates to increased risk according to the model (regression coefficient=1.89). On the other hand, the structural model between coping capacity and risk shows a negative coefficient, which means that increased coping capacity relates to decreased risk according to the model; regression coefficient=-0.16 (vehicle state) / -0.74 (operator state).

4.2 Harsh braking

Risk was measured through the types of harsh events and the level of severity at which they occur. For harsh braking, there are high severity events (1st STZ level), moderate severity events (2nd STZ level) and low severity events (3rd STZ level). Figure 3 illustrates the SEM results for harsh braking per each phase.

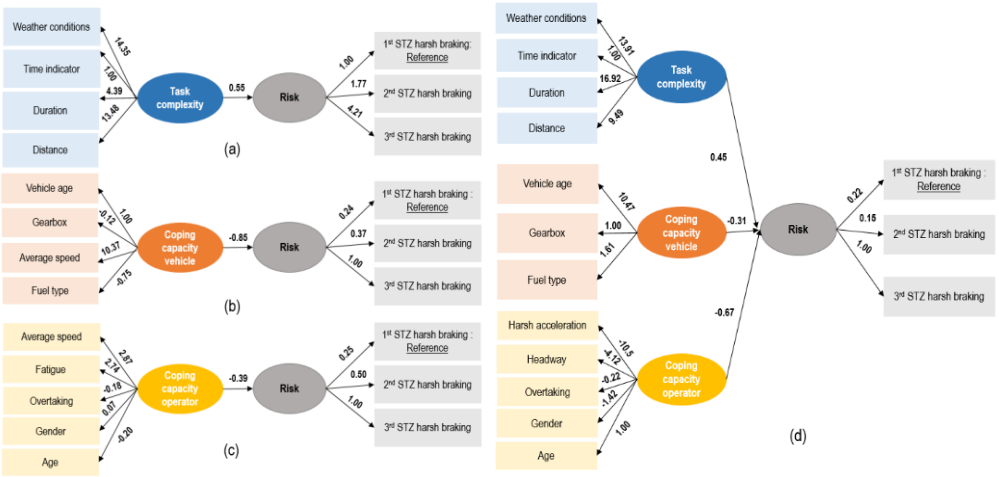

Figure 3. Results of SEM for harsh braking; experiment phase 1 (a), 2 (b), 3 (c), 4 (d)

Table 1. Model Fit Summary for speeding and harsh braking

Model fit metrics	Task complexity	Coping capacity operator state	Coping capacity vehicle state	Synthesis	Task complexity	Coping capacity operator state	Coping capacity vehicle state	Synthesis
	Speeding				***Harsh braking***			
p-value	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
CFI	0.997	0.992	0.730	0.952	0.981	0.707	0.897	0.832
TLI	0.994	0.987	0.601	0.936	0.969	0.613	0.848	0.845
SRMR	0.048	0.045	0.153	0.074	0.032	0.090	0.066	0.064
RMSEA	0.062	0.080	0.440	0.100	0.052	0.148	0.101	0.085

As shown in Table 1, p-values <0.001 across all metrics revealed strong evidence of model fit for both speeding and harsh braking. For speeding, all metrics demonstrated high levels of fit, with CFI values ranging from 0.997 to 0.981, TLI values ranging from 0.994 to 0.969, SRMR values between 0.032 and 0.066, and RMSEA values between 0.052 and 0.101. Similarly, for harsh braking, significant p-values alongside CFI values ranging from 0.992 to 0.707, TLI values between 0.987 and 0.613, SRMR values from 0.045 to 0.090, and RMSEA values spanning 0.080 to 0.148 affirm the robustness of the model fit across various dimensions, highlighting its reliability in capturing the complexities associated with both driving behaviors. The effect of the weather conditions was only considered in the "harsh braking" model and not the "speeding" model after continuous testing of multiple variables based on the optimum efficiency.

The task complexity latent variable takes into account all the variables and indicators obtained in the excessive speed category above, as well as the indicator of weather, which indicates the weather conditions prevailing on the route in question. It is noted that there is a positive correlation between trip complexity and risk (regression coefficient=0.55). The structural model between coping capacity and risk shows a negative coefficient, which means that increased coping capacity relates to decreased risk according to the model (regression coefficient=-0.85 (vehicle state) / -0.39 (operator state). Overall, there is a positive correlation between task complexity and risk (regression coefficient=0.45), while there is a negative correlation between vehicle and operator state and risk. On the other hand, SEM between coping capacity and risk shows a negative coefficient, which means that increased coping capacity relates to decreased risk (coefficient=-0.31 (vehicle state) / -0.67 (operator state)).

5 Conclusions

The ultimate goal of the analyses in this work was to identify the impact that the balance between task complexity and coping capacity has on the risk of a crash. To that end, 30 drivers participated in a naturalistic driving experiment carried out in UK and a large dataset of thousand of trips was analyzed. Explanatory variables of risk and the most reliable indicators, such as time headway, distance traveled, speed, time of the day (lighting indicators), or weather conditions were assessed.

To fulfill the aforementioned objective, SEMs were used to explore how the model variables were interrelated, allowing for both direct and in-direct relationships. Results demonstrated that coping capacity and risk displayed a negative relationship across all phases, indicating that higher coping capacity was associated with reduced risk.

The relationship among task complexity, coping capacity and risk, may depend on the specific context and the type of task or activity involved. In general, higher task complexity may increase the potential for errors or crashes, as it can lead to greater cognitive or physical demands on the individual performing the task. Similarly, a higher coping capacity may help to reduce the risk of crashes or errors, as it can provide individuals with the resources or strategies needed to effectively manage challenging or stressful situations. It is important to consider the specific factors and context involved when assessing the relationship among task complexity, coping capacity, and risk.

References

1. Ofori, E. K., Aram, S. A., Saalidong, B. M., Gyimah, J., Niyonzima, P., Mintah, C., & Ahakwa, I. (2023). Exploring new antecedent metrics for safety performance in Ghana's oil and gas industry using partial least squares structural equation modelling (PLS-SEM). Resources Policy, 81, 103368.
2. Useche, S. A., Hezaveh, A. M., Llamazares, F. J., & Cherry, C. (2021). Not gendered… but different from each other? A structural equation model for explaining risky road behaviors of female and male pedestrians. Accident Analysis & Prevention, 150, 105942.
3. Washington, S., Karlaftis, M., Mannering, F., & Anastasopoulos, P. (2020). Statistical and Econometric Methods for Transportation Data Analysis. Chapman and Hall/CRC.
4. World Health Organization - WHO (2018). Global status report on road safety 2018: summary (No. WHO/NMH/NVI/18.20). Last accessed on 25/5/2022. Retrieved from https://www.who.int/publications/i/item/9789241565684.
5. Chen, F. F. (2007). Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance. Structural Equation Modeling: A Multidisciplinary Journal, 14(3), 464–504.