Advancing simulation-based driver training

Joost de Winter

Advancing simulation-based driver training

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof. dr. ir. J.T. Fokkema,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op dinsdag 27 januari 2009 om 10.00 uur
door Joost Cornelis Franciscus DE WINTER
Ingenieur luchtvaart en ruimtevaart
geboren te Utrecht

Dit proefschrift is goedgekeurd door de promotoren:
Prof. dr. ir. P.A. Wieringa, Technische Universiteit Delft
Prof. dr. ir. J.A. Mulder, Technische Universiteit Delft

Samenstelling promotiecommissie:
Rector Magnificus, voorzitter
Prof. dr. ir. P.A. Wieringa, Technische Universiteit Delft, promotor
Prof. dr. ir. J.A. Mulder, Technische Universiteit Delft, promotor
Prof. dr. H.H. Bülthoff, Max Planck Institute for Biological Cybernetics, Duitsland
Prof. dr. K.A. Brookhuis, Rijksuniversiteit Groningen
Prof. dr. E. de Graaff, Technische Universiteit Delft en Aalborg University, Denemarken
Dr. ir. M. Mulder, Technische Universiteit Delft
R.W. Allen, M.S., Systems Technology, Inc., Verenigde Staten
Prof. dr. ir. J. Dankelman, Technische Universiteit Delft, reservelid

Dit onderzoek is gesubsidieerd door het Nederlands Ministerie van Economische Zaken, onder het innovatiegerichte onderzoeksprogramma Mens-Machine Interactie, IOP MMI. Titel project: Virtual Assistant.

Ir. J. Kuipers heeft in belangrijke mate bijgedragen aan de initiatie en ondersteuning van dit onderzoeksproject.

Ontwerp kaft: Jeroen den Dekker

ISBN/EAN: 978-90-9023347-5

Contents

Summary

Advancing simulation-based driver training

Road traffic crashes of young drivers are a major public health problem in all motorized countries. Research has shown that current on-road driver training is not effective in reducing these crashes, warranting the need for advancement. Driving simulators offer complementary advantages to on-road training: free control over training conditions, standardization, and objective driver assessment. Contemporary simulator developers do not fully exploit these possibilities; focus is often on the improvement of hardware and software in order to provide a realistic driving experience. However, it is unclear what level of realism (i.e., fidelity) is required for effective training. This is especially true for physical motion cueing: Although motion platforms have been useful for increasing user acceptance and for improving in-simulator performance, they are a major cost driver in the simulation industry and it is unknown whether they improve drivers' proficiency.

The first objective of this thesis is to exploit the driver assessment possibilities of simulators. In particular, the aim is to develop a method that can process raw measurement data into meaningful indicators about the learner driver, which can subsequently be used for student-adaptive feedback, instructions, and guidance. The second objective is to develop knowledge regarding how simulator fidelity – especially motion cueing – relates to training effectiveness.

First, this thesis presents a literature study on driver behaviour models and an analysis of experimental data in order to find out what model is suitable for constructing a student-profile. It is found that adequately understanding driver behaviour is not possible through a qualitative motivational model, nor can it be achieved with an adaptive control model that tries to describe what a driver exactly does at a certain moment. Instead, exploratory factor analysis is proposed; a statistical method for explaining individual differences. This method uses the matrix of correlations amongst diverse measures to describe these data by means of a small number of underlying factors.

Next, factor analysis is applied on data of a large number of participants who completed a driver-training programme in a simulator. From the task failures and the mean task completion times, three factors are extracted that are interpreted as errors, violations, and speed. Errors are unintentional, whereas violations represent intentional deviations from normal or recommended behaviour. Previous studies have used questionnaires for investigating the distinction between violations and errors. The present study is probably the first to extract the violation-error distinction from driving simulator data. Speed represents an individual characteristic that underlies the task completion times. Speed was found to have a positive correlation with violations but a negative correlation with errors. The validity of the three factors is assessed by investigating their relationships with gender, age, and the results of the on-road driving licence test. In accordance with literature about on-road driver behaviour, men were found to have a higher speed-score, a higher violation-score, and a lower error-score than women. Older participants had a lower violation-score and a lower speed-score than younger participants. Earlier licensure was statistically associated with a higher speed-score, a lower error-score, and a lower violation-score. It is recommended to employ the factor-scores in a student-profile which can be used for student-adaptive training.

Next, a literature study is provided on the relationship between driving simulator fidelity and training effectiveness. It is found that fidelity requirements are determined by a compromise, in which positive effects of an intervention (validity, transfer, and user acceptance) should be weighted against negative effects (cost, complexity, distraction, cue artefacts/conflicts). This thesis subsequently investigates whether low-cost motion cueing systems can satisfy this compromise of advantages and disadvantages. The following seven systems are experimentally tested in elementary braking and cornering tasks: a motion seat, a seatbelt tensioning system, a stiff brake pedal, a vibrating steering wheel, screeching tyre sound, a vibrating seat, and a pressure seat. The results show that most of these systems increase participants' ratings of realism, improve in-simulator performance, or result in lower vehicle decelerations. Hence, these systems satisfactorily fulfil the examined functions of motion platforms at a lower cost, thereby providing a good solution to the aforementioned compromise. Experiments are still needed that compare different motion systems regarding the transfer of training from the simulator to the roads.

A shortcoming of low-cost simulators is the limited amount of available sensors to record driver behaviour. Current driving simulators do not have eye-tracking systems and therefore cannot provide feedback on important visual tasks such as mirror-checking. Therefore, an experiment is conducted to investigate the effect of feedback on mirror-checking in a driving simulator. The results show that feedback led to improved learning for experienced drivers, but there was no benefit for inexperienced drivers. It is recommended to improve the didactic aspects of the simulator before augmenting the simulator with complex eye-tracking hardware.

This thesis concludes that factor analysis is a valuable method for constructing a student-profile on driving skill and driving style. The second conclusion is that low-cost motion cueing systems are valuable substitutes for more complex motion platforms. For future work, it is recommended to develop methods to suppress violating behaviour, to investigate transfer of training from the simulator to the roads, and to study the predictive validity of simulators regarding a person's accident proneness.

Joost de Winter

Samenvatting

Verbeteren van op simulatie gebaseerde rijtraining

Verkeersongevallen van jonge bestuurders vormen een groot volksgezondheidsprobleem in alle gemotoriseerde landen. Onderzoek heeft aangetoond dat de huidige rijtraining op de weg niet effectief is in het verminderen van deze ongevallen; er is behoefte aan verbetering. Rijsimulators bieden voordelen complementair aan training op de weg: vrije controle over de trainingscondities, standaardisatie en een objectieve beoordeling van de bestuurder. De ontwikkelaars van simulators maken nu niet optimaal gebruik van deze mogelijkheden; de focus ligt veelal op de verbetering van hardware en software met het doel een meer realistische rijervaring te creëren. Het is echter onduidelijk welke graad van realisme nodig is voor een effectieve training. Dit geldt vooral voor fysische bewegingsinformatie. Hoewel bewegingsplatforms nuttig zijn gebleken voor het verhogen van acceptatie onder gebruikers en voor het verbeteren van prestaties in de simulator, zijn ze duur en is het onduidelijk of ze de vaardigheid van bestuurders verbeteren.

Het eerste doel van dit proefschrift is het benutten van de mogelijkheden van prestatiebeoordeling van rijsimulators. Meer in het bijzonder is het doel een methode te ontwikkelen die rauwe meetgegevens kan verwerken tot betekenisvolle indicatoren over de leerling, die gebruikt kunnen worden voor leerling-adaptieve training, instructies en sturing. Het tweede doel is kennis te verkrijgen in de wijze waarop de graad van realisme van een simulator – bewegingsinformatie in het bijzonder – samenhangt met de effectiviteit van de training.

Eerst wordt in dit proefschrift een literatuuronderzoek uitgevoerd naar bestuurdersmodellen en worden experimentdata geanalyseerd om te achterhalen welk model geschikt is voor het opstellen van een leerling-profiel. Geconcludeerd wordt dat bestuurdersgedrag niet adequaat begrepen kan worden met behulp van een kwalitatief motivatiemodel en ook niet met een adaptief controlemodel dat tracht te beschrijven wat de bestuurder precies doet op een bepaald moment. Er wordt daarentegen een statistische methode geadviseerd voor het onderzoeken van individuele verschillen, namelijk exploratieve factoranalyse. Deze methode gebruikt de correlatiematrix van verschillende maten om deze data te verklaren door middel van een klein aantal onderliggende factoren.

Vervolgens wordt factoranalyse toegepast op de data van grote groepen deelnemers die een rijtrainingsprogramma hebben doorlopen in een rijsimulator. Uit de taakfouten en de gemiddelde duur van de taken worden drie factoren geëxtraheerd, die geïnterpreteerd zijn als fouten, overtredingen en snelheid. Fouten zijn onopzettelijk, terwijl overtredingen opzettelijke deviaties van normaal of geadviseerd gedrag voorstellen. Eerdere studies naar het onderscheid tussen fouten en overtredingen maakten gebruik van vragenlijsten. De huidige studie is waarschijnlijk de eerste die het onderscheid tussen fouten en overtredingen heeft geëxtraheerd uit simulatordata. Snelheid is een persoonskenmerk dat onderliggend is aan de duur van taken. De resultaten laten zien dat snelheid een positieve correlatie heeft met overtredingen, maar een negatieve correlatie met fouten. De validiteit van de drie factoren is beoordeeld door het onderzoeken van het verband met geslacht, met leeftijd en met de resultaten van het rijexamen op de weg. In overeenstemming met literatuur over rijgedrag op de weg, hadden mannen een hogere snelheidscore, een hogere overtredingscore, en een lagere foutscore dan vrouwen. Oudere deelnemers hadden een lagere overtredingscore en een lagere snelheidscore dan jongere deelnemers. Het eerder behalen van het rijbewijs hing statistisch samen met een hogere snelheidscore, een lagere foutscore en een lagere overtredingscore. Het wordt aanbevolen om de factorscores te implementeren in een leerling-profiel, dat gebruikt kan worden voor leerling-adaptieve training.

Dit proefschrift beschrijft vervolgens een literatuurstudie naar de relatie tussen de graad van realisme van een rijsimulator en de effectiviteit van de training. Geconcludeerd wordt dat de vereiste graad van realisme bepaald wordt door een compromis, waarin de positieve effecten van een interventie (validiteit, trainingsoverdracht, acceptatie van gebruikers) moeten worden afgewogen tegen de negatieve effecten (kosten, complexiteit, afleiding, onvolkomenheden/conflicten in de aangeboden informatie). Hierna is onderzocht of goedkope systemen die bewegingsystemen aanbieden kunnen voldoen aan dit compromis van voor- en nadelen. De volgende zeven systemen zijn experimenteel getoetst tijdens elementaire rem- en stuurtaken: een bewegende stoel, een aanspannende riem, een stijf rempedaal, een vibrerend stuurwiel, piepend bandengeluid, een trillende stoel en een drukstoel. De resultaten geven aan dat de meeste van deze systemen zorgen voor een door de deelnemers hoger ingeschat realisme, voor verbeterde prestaties in de simulator of voor kleinere remvertragingen. Ofwel, deze systemen vervullen de onderzochte functies van bewegingsplatforms naar tevredenheid tegen lagere kosten en vormen daarmee een goede oplossing voor bovengenoemd compromis. Er is nog behoefte aan experimenten die een vergelijk maken tussen simulators met verschillende bewegingssystemen met betrekking tot de overdracht van training van de simulator naar de weg.

Een tekortkoming van goedkopere simulators is dat daarin een beperkt aantal sensors aanwezig is om het bestuurdersgedrag te registreren. De huidige simulators hebben geen kijk(richting)registratieapparatuur en kunnen daarom geen feedback geven op belangrijke visuele taken zoals het controleren van de spiegels. Daarom is een experiment uitgevoerd dat de effecten van feedback op het gebruik van de spiegels onderzoekt in een rijsimulator. De resultaten geven aan dat feedback leidt tot verhoogde leerprestaties bij ervaren bestuurders, maar er was geen voordeel voor onervaren bestuurders. Het wordt aanbevolen om de didactische aspecten van de simulator te verbeteren, alvorens de simulator te voorzien van complexe kijkregistratieapparatuur.

Dit proefschrift concludeert dat factoranalyse een waardevolle methode is voor het opstellen van een leerling-profiel over rijprestatie en rijstijl. De tweede conclusie is dat goedkope systemen die bewegingsinformatie aanbieden nuttige alternatieven zijn voor complexere bewegingsplatforms. Voor toekomstig onderzoek wordt aanbevolen om methoden te ontwikkelen die overtredinggedrag kunnen onderdrukken, te onderzoeken of de vaardigheden die geleerd zijn in de simulator overdraagbaar zijn naar de weg, en de voorspellende waarde van simulators te onderzoeken met betrekking tot iemands kans op een verkeersongeval.

Joost de Winter

CHAPTER 1

Introduction: Issues in on-road driver training and prospects for simulation-based training

De Winter, J.C.F., De Groot, S., Mulder, M., & Wieringa, P.A. (2008b). *Issues in on-road driver training and prospects for simulation-based training*. Parts of this manuscript have been submitted for consideration in the Journal of Safety Research, Elsevier.

1. Introduction

1.1. Young driver problem

Road traffic crashes are a major health problem in all motorized countries. Worldwide, approximately 1.2 million fatalities occur in road traffic every year and millions more get injured or disabled. The high-income countries account for about 10–15% of the deaths and annual disability-adjusted life years (Peden et al., 2004). If the European Union countries were able to prevent all road traffic crashes and associated costs, a 162 billion Euro socioeconomic benefit would arise annually (European Transport Safety Council, 2003), about 2% of the gross domestic product.

Figure 1 shows the number of victims of car crashes in the Netherlands with a severity of injury that led to hospitalization. It can be seen that younger people, particularly men, are overrepresented. The overrepresentation of young drivers in car crashes is also referred to as the *young driver problem* (Organisation for Economic Co-operation and Development [OECD], 2006). In the OECD countries, fatalities in road traffic per million population occur twice as frequently among drivers who are younger than 25 years as compared to older drivers (OECD, 2006). Globally, road traffic crashes are the leading cause of death among 15–19-year-olds (Toroyan & Peden, 2007). Figure 1 also shows that, during the last decades, the number of crashes decreased considerably. Because this decrease was more pro-

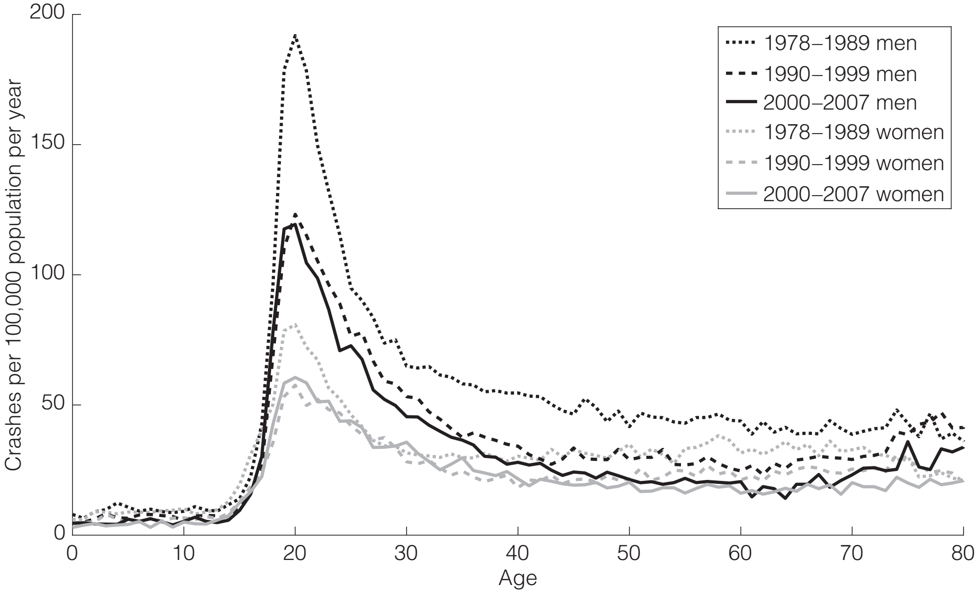


Figure 1. Number of police-registered victims of car crashes in the Netherlands with a severity of injury that led to hospitalization, as a function of age, gender, and year (SWOV, 2008).

nounced for older persons, the crash involvement of younger persons has become *relatively* more prominent. Based on Figure 1, when comparing the period 1990–1999 with the period 2000–2007, the crash risk of the 26–80 age group declined with 22%, whereas the crash risk of the 16–25 age group declined with only 1%. Similar trends, based on fatal crash data per kilometre travelled, have been observed in other countries as well (OECD, 2006; Twisk & Stacey, 2007). A possible explanation for the relative increase of the young driver problem is that generic safety measures, such as improved crashworthiness of vehicles and better enforcement, had positive effects for particularly the older drivers, whereas measures that were specifically aimed at young drivers, such as changes in driver education, had no effect on crash statistics (Vlakveld, 2006a).

As a simplification, it is assumed that the three main contributing factors to the young driver problem are age, inexperience, and gender (OECD, 2006). With regard to age: Younger drivers have a higher tendency for sensation seeking behaviour and unsafe lifestyles than older drivers (Arnett, 1996; OECD, 2006). The second contributing factor, interacting with age, is inexperience. Learning to drive needs extensive practice to reach a sufficient level of skill (OECD, 2006). As a novice driver gains experience, mental models are formed, perceptual skills improve, and fewer mental resources are required for executing driving tasks correctly (e.g., Drummond, 1989; Nyberg, 2007). The third contributing factor is gender. Young men drive more than young women, and have higher tendencies for risk factors such as fast driving and violating traffic rules (OECD, 2006). In addition to these three factors, specific risk factors have been identified, such as distraction by infotainment systems (J.D. Lee, 2007) and sleep deprivation (Groeger, 2006).

The interacting effect of age and experience is illustrated in Figure 2 (adapted from Maycock & Lockwood, 1993). It can be seen that age has a positive effect on crash risk, but driving experience has a stronger influence. Crash risk decreases sharply within the first few months (or thousands of kilometres) after passing the driving test (see also Emmerson, 2008; Mayhew et al., 2003; McKnight & McKnight, 2003; Vlakveld, 2005a). In other words, although newly licensed drivers have experience in the form of driving lessons, they start driving independently at a safety level considerably lower than that reached after only a few months of independent driving. Major safety benefits would result if the learning now occurring during the first months of independent driving could be achieved under protected training conditions.

1.2. The role of simulation-based training

Currently, an important trend in driver training is the increased use of technological aids. A number of CD-ROM-based training packages have been developed (Senserrick & Haworth, 2005), and driver-training simulators are becoming commonplace (e.g., Allen et al., 2007a; Welles & Holdsworth, 2000). In the EU-funded

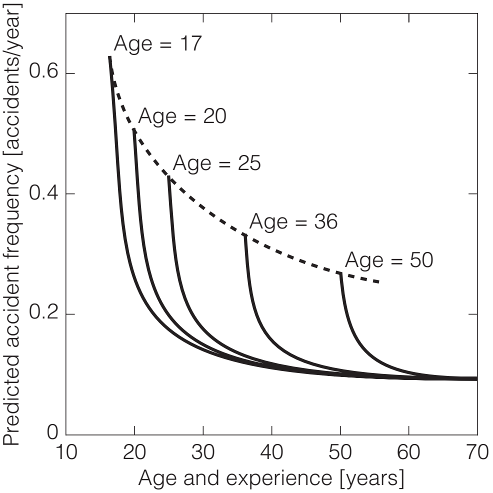


Figure 2. The predicted effect of age and driving experience on accident liability for drivers whose annual mileage is 7,500. The figure is constructed from a model based on questionnaire responses of 13,519 drivers. The dashed line represents the effect of age on accident liability. The solid lines represent the effects of experience for drivers who start to drive at the ages shown in the figure.

Adapted from Maycock, G., & Lockwood, C.R. (1993). The accident liability of British car drivers. *Transport Reviews, 13*, 231–245, Figure 2. Reuse offered by Taylor & Francis.

TRAINER project, research was conducted into the use of driving simulators as a means to improve road safety (Dols et al., 2001). Another EU-funded project, called TRAIN-ALL, aims to develop a cost-effective driver training and assessment system (Panou & Bekiaris, 2007). The increasing role of simulators in driver training is a phenomenon that warrants further investigation.

In this chapter, we provide an overview of the effectiveness of current on-road driver training with regards to road safety. Moreover, we explore the potential role of simulation-based driver training to improve training effectiveness. We show which issues are to be tackled in on-road driver training and what opportunities and challenges lie ahead in simulation-based training. Finally, we make recommendations regarding research needed to improve training effectiveness, to the benefit of safety on the road.

2. On the effectiveness of on-road driver training

The original *Holy Trinity* of traffic safety measures comprised the three E's: Education, Enforcement, and Engineering. Today, many interventions of the latter two E's have shown to be effective, such as improved crashworthiness of cars, guardrails, yield or stop signs at intersections, speed humps, mini roundabouts, setting and enforcing speed limits, seatbelt legislation and enforcement, as well as the introduction and enforcement of laws on blood alcohol concentration (Elvik & Vaa, 2004; Toroyan & Peden, 2007). However, education in the form of formal driver training is generally not in the list of effective crash countermeasures.

2.1. The effectiveness of formal driver training

The primary goal of driver training and testing should be to ensure that new road users drive safely. A meta-analysis of Elvik and Vaa (2004) concluded, however, that formal driver training is not an effective road safety measure. The analysis included 16 studies that compared formal driver training by driving schools with informal driver training, that is, self-training or training provided by family or friends. After selecting the methodologically best studies (i.e., experiments that distributed participants randomly between formal and informal driver training), results showed that formal driver training resulted in 0% difference in the number of crashes per driver and 11% more accidents per kilometre driven as compared to informal training. In addition, experiments showed that the more lessons one had taken, the more the crash rate increased (Elvik & Vaa, 2004). One may argue against the validity of the meta-analysis of Elvik and Vaa (2004) as it comprised many older studies, such as the large DeKalb study from the 1980s (see Lund et al., 1986). Psychological processes that are currently known to contribute to the ineffectiveness of formal driving training (see section 2.5) were not used in designing these older curricula. In addition, many studies had been carried out in the United States and (oppositely to European curricula) were characterized by lots of classroom education and little on-road training (see Vlakveld, 2006a, for a discussion).

Several research papers are more optimistic about driver training. For example, a study from Denmark stated that nationwide changes of the training curriculum in 1986, including incorporation of hazard perception and defensive driving, were (partially) responsible for a decrease in crash risk (Carstensen, 2002). Unfortunately, that study, as many others, was not a controlled trial and therefore methodologically weak. According to Elvik and Vaa (2004), the results of their meta-analysis cannot be explained because poorer training schemes were evaluated. It was far more probable that the evaluated studies were relatively well thought-out programmes. The results cannot be explained by the fact that the research was of poor quality either. On the contrary, there was a tendency that the methodologically better studies yielded less favourable effects on road safety.

Various other overviews exist showing that formal driver training is ineffective (e.g., Brown, 1997; Christie, 2001; Mayhew et al., 1998). Scientific reviews and meta-analyses have not been optimistic on the safety-benefits of the post-licence Defensive Driving Course in the United States (Lund & Williams, 1985), high school driver education (Vernick et al., 1999), and post-licence driver education programmes

(Ker et al., 2005). Summarized, current formal training programmes appear to offer no benefits in terms of safe driving as compared to informal training.

2.2. The effectiveness of other types of training

Elvik and Vaa (2004) also meta-analysed the effects of knowledge training, special skills training, and training of certain groups of drivers. No clear statistical relationship between driver's theoretical knowledge and crash rates was found. Skid training and night driving courses significantly increased the number of crashes, which is remarkable, considering that the majority of such courses intend to teach skills in avoiding crashes, such as evasive manoeuvring. Courses for older drivers did not appear to affect the number of crashes per driver either. However, formal training of professional drivers, in particular training in defensive driving taught at the workplace combined with motivation and incentive systems, did reduce crash rate. Other experimental studies showed that teaching defensive driving to drivers who had been previously convicted for traffic offences reduced the number of crashes (see also Masten & Peck, 2004, for a meta-analysis demonstrating a positive effect; however, Ker et al., 2005, for a meta-analysis describing that there is no effect). To summarize, the effects of other types of driver training are rather mixed. Training of technical driving skills in some cases increases crash risk, whereas safety benefits are observed in some intensive attitude related training programmes.

2.3. The effectiveness of accompanied practice

As mentioned, inexperience is a key contributing factor to the young driver problem. Therefore, recent guidelines recommend high levels or extended periods of pre-licence accompanied practice (referred to as lay instruction, e.g., with parents) (Hatakka et al., 2003; OECD, 2006). At present, accompanied practice is allowed in 15 of the 27 European Union countries, but few countries actively encourage practice to increase novices' experience by the time they start driving independently (Twisk & Stacey, 2007). Positive associations between accompanied practice and road safety have been found in Sweden, Finland, and Austria (Gregersen et al., 2000; Twisk & Stacey, 2007); mixed results were found in France and Norway (Twisk & Stacey, 2007). In all cases, causal relationships were not evident as they were no randomized controlled trials. That is, those drivers who chose to drive with parents could be inherently safer than those who did not choose to do so.

Although researchers' opinions about accompanied practice are often positive, there exist important drawbacks that should be mentioned. Disadvantages are that, firstly, not everyone has the possibility of doing so. Low popularity and stress between parents and teens can be problematic as well (Simons-Morton & Ouimet, 2006; Twisk & Stacey, 2007). Second, there is lack of standardization. Focus is on learning by doing, whereas goals, contents, and structure of training are subject to wide variability (Hatakka et al., 2003). This may lead to the acquisition of bad habits. In France, the lack of success of accompanied practice was attributed to the fact that the trips undertaken were more standard (e.g., shopping and holidays), whereas for more demanding tasks the responsibility was delegated to the supervisor (Page et al., 2004; Twisk & Stacey, 2007). Third, although accompanied practice has been reported to be as much as 10–35 times safer than independent driving after obtaining a licence, it is not completely safe. Research from Sweden has shown that about three fatalities occurred per year during supervised practice (Gregersen et al., 2003). Fourth, a number of studies exist that have cast doubts on the effectiveness of accompanied practice (Simons-Morton & Ouimet, 2006). Groeger and Banks (2007) believed that there is little foundation for the hypothesis that what is learned under protective conditions (either accompanied driving or restrictions such as a nighttime curfew during the intermediate phase of graduated driving licensing) will have a safety benefit in later driving. Although crash risk drops when the protective measures are applied, drivers remain at risk when the restrictions are eased (Groeger & Banks, 2007; Males, 2007). To summarize, it is uncertain whether pre-licence accompanied practice is an effective remedy to the young driver problem. More research is needed.

2.4. Does formal training help to pass the driving test?

The most important reason why novices want to learn to drive is likely to obtain a driving licence, not to drive safely per se. Although there is substantial evidence that driving skills improve during formal training (e.g., Groeger & Clegg, 2007), it is questionable whether formal training is more effective than accompanied practice (per hour of driving) in letting students pass the driving test. Hatakka et al. (2003) concluded that pass rates tend to be higher in professional training than in more structured training models. However, causal effects are not evident. A study of Forsyth (1992) showed that, for more than 10 hours of training, those students who had received more hours of professional training had a lower chance of passing the driving test. Similarly, Hall and West (1996) found that instructors' ratings of the chance of passing the driving test in the first attempt was positively associated with prior hours of practice and negatively associated with prior formal tuition. In these studies, the direction of causality was difficult to ascertain. Groeger (2001) concluded that it is likely that worse drivers gradually gravitate towards professional instructors, but that the amount or nature of the tuition they then receive is not sufficient to enable them to improve as much as they need to.

2.5. Causes of the ineffectiveness of on-road driver training

It was shown that the effectiveness of current on-road driver training is rather poor, and at best, mixed. Numerous scientific studies have investigated the causes of the ineffectiveness of on-road driver training. Literature provides several possible interrelated reasons why formal driver training on the road does not improve traffic safety, which are described below.

2.5.1. Drivers adapt

Elvik and Vaa (2004) considered that drivers adapt their behaviour according to their perceived ability. Training may lead to earlier licensure and may create (over)confident drivers who are tempted to drive less carefully. Indeed, many studies stated that crashes are particularly related to poor safety strategies and committing of violations, not to poor basic vehicle handling skills per se (e.g., Clarke et al., 2005; Parker, 2007).

2.5.2. Safety is ignored in training

Another popular explanation is that current training is essentially guided to obtaining the driving licence as quickly as possible, and that it ignores much of the safety-relevant behaviour. Amongst others, Hatakka et al. (2003) recommended that driver training should put more emphasis on higher order skills such as risk-assessment, self-assessment, and the development of a responsible attitude.

2.5.3. Necessarily experience-based

Some have suggested that there is no foundation for the assumption that what is learned under restricted conditions will have a safety benefit (e.g., Groeger & Banks, 2007). Harrison (1999) asserted that the development of automaticity and internal models is necessarily experience-based and that it is difficult to conceive a training method which could replace the need for experience. Harrison noted that it might be possible to use a driving simulator for this purpose but recommended that more research into simulator fidelity, that is, its capability to provide a realistic driving experience, is required.

2.5.4. Psychologically unsound

Brown (1997) suggested that driving instruction might be psychologically unsound. Explanations were sought in the effectiveness of part-task training versus whole-task training, or distributed practice (short driving lessons) versus massed practice (long driving sessions). In addition, it may be difficult to learn whilst simultaneously performing a task and receiving declarative information from a human instructor. Furthermore, driving under supervision may prevent learning of autonomous decision-making and may make pupils anxious compared to driving with nonprofessionals. Drummond (1989) criticized the limited sophistication of course content. Courses covered a wide range of areas, but educational objectives were based neither on empirical evidence nor did they have a substantive theoretical foundation.

2.5.5. Too little training

Simons-Morton and Ouimet (2006) considered that an important cause of the ineffectiveness of current driver training in the United States is that far too few hours of professional road instruction were applied. Similarly, Brown (1997) considered that there are too few opportunities to practise and automate behaviour. A recent report from the United Kingdom provided disconcerting conclusions about the lack of varied training: “From a road safety perspective, it is worrying that one in twenty (5%) respondents who passed the test had no experience on country roads, given that a high proportion of casualties occur on rural roads. One in eight (12%) respondents who actually passed their practical test admitted that they had no experience at all of driving in darkness” (Emmerson, 2008, p. 11).

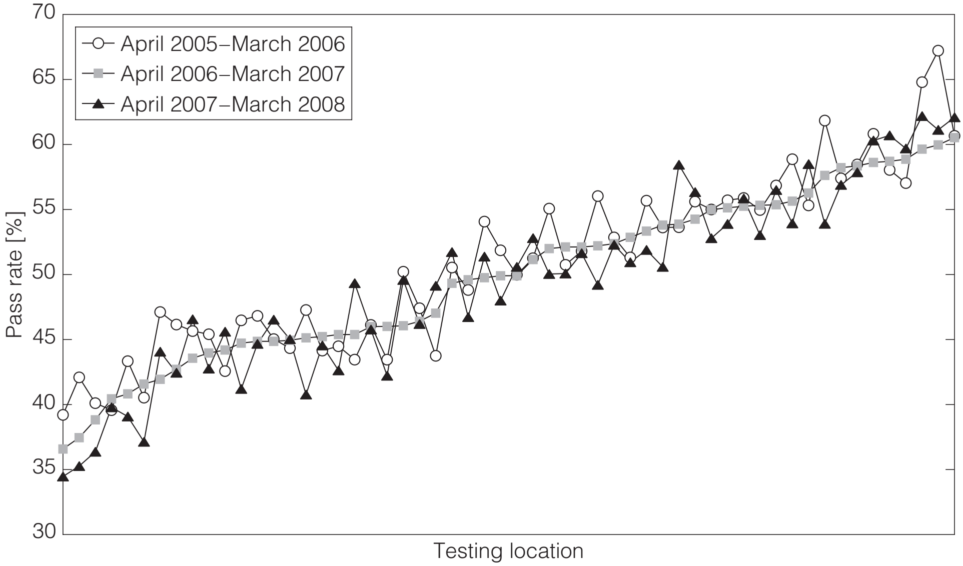


Figure 3. Pass rates on the first attempt on the official Dutch driving licence test as a function of testing location. The 56 regions are sorted on the pass rate in the period April 2006–March 2007. The lowest pass rates were observed in the more densely populated regions (e.g., Amsterdam, Rotterdam). The pass rates were approximated using publicly available data of all driving schools in the Netherlands (CBR, 2008). Because the pass rates and the number of students of driving schools were listed as categorical data (e.g., 0–5%, 6–10%, …), the most probable number for each category was used in the approximation.

Performance assessment during on-road driver training is inevitably subjective.

2.5.6. No objective data

Fairness and reliability are crucial aspects in psychological education and testing but are compromised during on-road training. Human performance assessments and corresponding feedback are inevitably subjective and sensitive to individual differences. Indisputably, subjective measures often contain valuable information that cannot be captured by objective performance measures. However, using a combination of objective and subjective measures can reveal more about task performance than a measure taken in isolation (Brookhuis & De Waard, 2002).

Groeger and Rothengatter (1998) described that feedback and error-correction procedures were not systematically applied, which may be one of the causes of the poor results of formal driver training. The discrepancy between human performance assessments (including self-assessments) and actual driving behaviour has been discussed extensively in the literature (e.g., Groeger, 2000a; Hatakka et al., 1997; Sundström, 2008). Baughan et al. (2005) found that there were substantial differences in the durations of the driving tests, addressed items, and pass rates between countries and test centres. Conditions on the road as well as driving test results were largely random. Figure 3 shows the approximated pass rates on the official driving test for the 56 testing regions in the Netherlands. It can be seen that fairly large and consistent differences in pass rates existed between testing locations. The causes of this unevenness cannot be inferred from these data. Possible explanations are regional differences in road infrastructure, amount and quality of driving lessons, or demographics. Nonetheless, these results do suggest that training or testing conditions differ across the country.

2.6. New perspectives

The rather pessimistic state-of-affairs of current driver training methods does not imply that it is impossible to create a more effective training strategy. Extensive research is undertaken to improve driver-training models. Many underscore the importance of research in traffic safety and driver training (e.g., Johnston, 1992; Mayhew, 2007; OECD, 2006; Peden et al., 2004). In a recent report (European Commission, 2007), a list of best practices was provided for improving driver training. The EU-project BASIC (Hatakka et al., 2003) as well as Williams (2006) recommended that diverse training methods need to be applied together because one method alone is not effective enough.

3. Simulation-based driver training

The use of driving simulators for training purposes is on the rise. At present, about 100 driving simulators are operational in the Netherlands for basic driver training (Kappé & Van Emmerik, 2005). The present section elaborates on past research, advantages, disadvantages, and unexploited opportunities of simulation-based driver training. Note that the essentials of training in driving simulators, such as the characteristics of simulator software and hardware, are introduced elsewhere (e.g., Kappé & Van Emmerik, 2005; SWOV, 2006).

3.1. Is simulator training effective?

There are several indications but no compelling evidence that simulator training speeds up skill acquisition of unlicensed drivers as compared to on-road training (Kappé & Van Emmerik, 2005; Vlakveld, 2006b). A study has shown that it is possible – at least for some individuals – to pass the driving licence test after just 9 hours simulator training in combination with about 30 minutes on-road practice (Wierda, 1996; described in Vlakveld, 2005b). In addition, studies have demonstrated transfer of training from the simulator to the road. Roenker et al. (2003) found that simulator training enhanced performance of at-risk older adults on particular driving tasks, although some of these gains disappeared over time. Simulator training showed transfer to the road of fuel-efficient driving (Strayer & Drews, 2003; Strayer et al., 2005) and manoeuvring of a truck (e.g., Uhr et al., 2003).

Few studies have investigated the effects of simulator training on drivers' crash involvement. An early study from 1973 (Jones, 1973; described in Elvik & Vaa, 2004) found that there was no significant difference in crash rate between drivers who had training in a simulator and drivers who were trained in regular traffic. Experiments found positive effects of computer-based training on hazard awareness and attention control in the simulator and on the road (e.g., Diete, 2008; Fisher et al., 2006; Senserrick & Haworth, 2005, for an overview). It has also been found that the results of computerized hazard perception tests were predictive of drivers' crash involvement on the road (Congdon, 1999; Vlakveld, 2008; Wells et al., 2008). Recently, Allen et al. (2007a) found that those individuals who had completed a simulator training programme involving repeated exposure to critical hazards in a wide field of view instrumented cab, had a post-licence crash risk that was only one third of that of the general teen population. However, the study was not a randomized controlled trial and a causal relationship was therefore not established. The only randomized controlled trail we could find that evaluated the effect of simulator training on crash risk was conducted by Strayer et al. (2005, 2008). These authors evaluated a purpose-developed training programme involving 4 hours of simulation for snowplough operators. The simulator curriculum incorporated part-task training and variable priority training methods. Results showed that in the 6-month period following training, the chance of being involved in an accident were lower for the drivers who had received training as compared to a matched control group that had not received training. Moreover, user ratings were high, with drivers indicating that the training helped them prepare to drive safely. For future research, it would be interesting to compare the effectiveness of simulation-based training to traditional training with a human instructor.

To summarize, there exists no evidence that pre-licence simulator training is more (or less) effective than on-road training. However, for professional snowplough drivers, it has been demonstrated that simulator training can reduce future crash risk. Finally, it can be said that simulation-based training is relatively new and unstudied compared to driver training on the roads.

3.2. Advantages of driving simulation

Driving simulators offer advantages and complementary possibilities to formal driver training on the road. These include (a) operational advantages, (b) control over training conditions and standardization, (c) novel possibilities for feedback and instruction, and (d) objective performance measurement.

3.2.1. Operational advantages

Simulators can prepare drivers to handle unpredictable or safety-critical tasks that are inappropriate to practise on the road, such as collision avoidance or risky driving (Hoeschen et al., 2001). Making mistakes may be a key dimension to learning. Flach et al. (2008) stated: "This is likely to be one of the values of simulators – they offer an opportunity to learn from mistakes in a forgiving environment" (p. 134). Allen et al. (2007a) made similar remarks: "Motor vehicle crashes are significantly higher among young drivers during the first year of licensure, and crash risks decline with increased experience …. This produces an interesting dilemma about how to provide young drivers with driving experience without significantly increasing their crash risk. Driving simulation may be the solution to this dilemma, since exposure to hazardous driving conditions can be simulated in a controlled and repetitive way without risk." The sheltered conditions in a simulator provide another desired effect. Results of interviews with simulator students and driving school owners indicated that reduced nervousness was regarded as one of the primary advantages to start training in a simulator instead of a real car (Van der Snee, 2005).

Training in a simulator can be cheaper than training in the operational environment. In flight simulation, the corresponding cost ratios between real training and simulation-based training far exceed 1:20 (see Verstegen, 2003). Obviously, these cost ratios will be less beneficial for a task as car driving (Wheeler & Trigs, 1996). Nonetheless, driving simulators are being used for driver training on a cost-effective basis (e.g., Kappé & Van Emmerik, 2005). One of the main cost-savings of using a simulator stems from the fact that – with a well-developed intelligent tutoring system – a human driving instructor in the traditional sense is not needed anymore (Kappé et al., 2003; Weevers et al., 2003b). Then again, not all simulators are cost-effective devices. It has been suggested that the investments that were required for certain (research) driving simulators may have been unjustified (Evans, 2004). Finally, simulators do not consume fuel and do not use the road infrastructure. Although quantitative data are unavailable, it is likely that fixed-base simulators use less energy and are responsible for less CO_2 emissions than a real car.

3.2.2. Control over training conditions and standardization

Simulators provide complete control over training conditions. Behaviour of other vehicles, weather conditions, or the virtual environment can be manipulated in real time according to the training needs (Wassink et al., 2006). It is also possible to confront a trainee with novel technical devices such as adaptive cruise control, front wheel drive or rear wheel drive, or vehicles of different masses, which can make them aware that they should adapt their behaviour (Hoeschen et al., 2001). Furthermore, it is possible to (partially) automate driving tasks. For trainees who practise steering, use of throttle and gear changing can be assisted or automated, reducing workload and allowing more attention allocated to the main training task. Virtual environments are purpose-developed as well, making it possible to practise many manoeuvres in a short training session. In addition, simulators offer the possibility of standardization of training conditions. Students at different locations can drive under the exact same conditions, if the simulators have identical hardware and software.

3.2.3. Novel possibilities for feedback and instruction

De Groot et al. (2007) provided an overview of the didactic possibilities of driving simulators. Simulators can show demonstrations, replays, and video instructions to present courseware of real-world scenarios. Simulators allow for the possibility of performing part-task exercises. Finally, simulators have the possibility to provide feedback and instructions in modalities other than speech, such as visual interfaces, auditory signals, tactile feedback (such as a vibrating seat), or augmented cueing and feedback (Hoeschen et al., 2001; Van Emmerik, 2004). A recent study found beneficial effects from multimodal instruction in driving simulation. Here, visual route instructions were provided complementarily to traditional preprogrammed route instructions, diminishing the number of driving errors (De Groot et al., 2006).

3.2.4. Objective performance measurement

A driving simulator can measure performance automatically, objectively, and accurately. Multiple channels (e.g., steering wheel angle, vehicle speed, intervehicle spacing) can be recorded and stored in memory. As this PhD thesis will demonstrate, these data can be used for objective diagnosis of student performance, statistical analyses, and predictions of future driving performance. Objective measurement and data storage is hardly possible during on-road training, unless an instrumented vehicle is used. McCall et al. (2004) provided insight into the significant amount of technological systems that are needed to record a real vehicle's state and its near surroundings.

3.3. Disadvantages of driving simulation

Most criticism on driving simulators boils down to their limited realism. Per definition, driving in a simulator is not the same as driving in the real world. In other words, simulators have limited fidelity. The following fidelity-related disadvantages of driving simulation have been identified:

1. It has been argued that particular skills, such as vehicle manoeuvring, cannot be properly trained in a simulator because of the discrepancy between the cues offered by the simulator and the cues offered in reality (Kappé & Van Emmerik, 2005). More generally, insufficient transfer from what has been learned in the simulator to later activities is a concern of all simulation-based training.
2. Simulator sickness symptoms may undermine training effectiveness (see e.g., Mollenhauer, 2004).
3. Students may be less motivated by a limited-fidelity simulator and prefer a real vehicle instead. That is, there is the issue of user acceptance.
4. There is a lack of social context with other road users, which makes it difficult to train informal rules and human interaction (Kappé & Van Emmerik, 2005).
5. Real danger and real consequences of actions do not emerge (Käppler, 1993).
6. Learning more or less incidentally in varied real environments may leave an impression considerably longer than learning in a virtual environment (SWOV, 2006).

A possible downside of highly repetitive and structured training is that, although one may quickly learn a task, one may also quickly forget what has been learned. Nevertheless, a recent study has shown that part-task simulator training did in fact reduce drivers' future crash involvement (Strayer et al., 2008).

7. A computer can assess only performance on simple tasks. A human instructor is needed for more complex assessments such as performance in the appropriate use of the mirrors (Kappé & Van Emmerik, 2005).
8. Simulators are effective in assessing driving skills. However, road safety is primarily determined by driving style, what a driver chooses to do (Evans, 2004). Evans: "It is exceedingly unlikely that a driver simulator can provide useful information on a driver's tendency to speed, drive while intoxicated, run red lights, pay attention to non-driving distractions, or not fasten a safety belt. Twenty-year-olds perform nearly all tasks on simulators better than the 50-year-olds, but it is the 50-year-olds who have sharply lower crash risks" (p. 188).

Other disadvantages of simulators are that they can be technically complex and expensive, that they may require support facilities for editing software, and that developing a simulator may be a time-consuming process (Jamson et al., 2007; Verstegen, 2003). These issues particularly apply to high-fidelity simulation. Low- or medium-fidelity simulators, on the other hand, are often less complex and less expensive than a real vehicle.

3.4. Fidelity: a complicated matter

Self-evidently, training will be ineffective when the simulation deviates too much from reality. A seemingly rational method is therefore to replicate the operational environment to a high accuracy. Indeed, many researchers and developers spend considerable effort doing so, for instance, by introducing progressively sophisticated visual display systems, motion actuators, and motion cueing algorithms. However, research has shown that a highly realistic simulator is not desirable per se (e.g., Alessi, 2004; Salas et al., 1998). High-fidelity simulation of cues that are irrelevant to learning and have no functional purpose, such as detailed scenery, could be a misuse of resources or even detrimental to training effectiveness.

Physical motion constitutes a discrepancy between simulation and on-road driving that is especially worth mentioning. The accelerations that are provided by a six degree-of-freedom platform (during a stopping manoeuvre) do not accurately resemble those in the real vehicle (Tomaske et al., 2001), which may justify the use of even more sophisticated solutions such as centrifugal devices or gliding rails. Motion platforms have been widely used in flight simulation for years to train professional pilots. Although motion generally improves in-simulator performance and increases user acceptance, no objective evidence exists on whether simulator motion positively contributes to performance in the operational environment (Bürki-Cohen et al., 1998; McCauley, 2006). The need for further research into simulator motion has been expressed (e.g., Bürki-Cohen et al., 1998; Hays et al., 1992). A relevant research question remains whether fully realistic simulation of physical motion is needed. Perhaps much cheaper solutions, such as vibrations or limited-amplitude cueing, can be equally or even more valuable.

Several purported disadvantages are of paradoxical nature. For example, the reported disadvantage that real danger does not exist in simulation has been mentioned as an advantage as well. Interestingly, Turpin et al. (2007) found that the occurrence of simulator discomfort during training of police officers was actually less than the occurrence of dropout due to discomfort reported while driving real vehicles on the testtrack. In addition, although a simulator's ability to show a replay of driving performance is not feasible on the road and therefore not realistic, it could be effective for training.

Research into simulator fidelity is highly scattered, as also evidenced by the many expressions encountered in literature, including (but not limited to) physical fidelity, objective fidelity, perceptual fidelity, behavioural fidelity, functional fidelity, psychological fidelity, absolute fidelity, relative fidelity, statistical fidelity, and selective fidelity. It has been suggested that effective training can be acquired with low/medium-fidelity driving simulators (Allen et al., 2007a; Welles & Holdsworth, 2000), whereas others assert that high fidelity is a prerequisite (e.g., Harrison, 1999). A review on human perception in driving simulation indicated that past experiments have provided substantial insights, yet many questions remain unanswered about how simulator characteristics affect driving behaviour (Kemeny & Panerai, 2003). Blana (1996) and Kaptein et al. (1996) provide overviews of driving simulator fidelity.

To summarize, there are many unknowns regarding how simulator fidelity – and motion cueing in particular – affects training effectiveness.

3.5. Unexploited opportunities

As pointed out by various sources, current practice in research and development of simulation-based (driver) training is centred too much towards hardware specifications, and the real potential of simulators is not fully used (De Groot et al., 2007; Salas et al., 1998; Verstegen, 2003). Van Emmerik (2004) described a distinction between the traditional design perspective and the training perspective: "The traditional simulator design viewpoint is very much oriented towards fidelity …. they [its adherents] consider the characteristics of the hardware and the mathematical models to be the main determinants of the simulator's potential to make training effective and efficient …. The risk that looms for this approach is the neglect of instructional factors" (p. 10). Salas et al. (1998) noticed an identical problem in flight simulation training and described the distinction between the interests of engineers and computer scientists on the one hand, and psychologists and training specialists on the other. According to Salas et al. (1998), instructional features, such as feedback, performance measurement, and scenario design are important aspects of learning but not systematically applied.

Verstegen (2003) came to similar conclusions after conducting research on 44 relatively high-fidelity training simulators, mostly on military training sites. The study evaluated simulators of vehicles such as helicopters and ground vehicles as well as communication and navigation systems. Several shortcomings of these simulators were described. For example, less than 10% of the simulators featured automatic feedback or scoring, implying that the amount and quality of feedback depended very much on the instructor's expertise and experience (Verstegen, 2003). Only one simulator featured facilities for long-term data storage. Verstegen: "The lack of facilities to support feedback is remarkable …. Frequently occurring errors in procedural tasks are often easily detectable or measurable. In these cases, the simulators could provide feedback automatically …. The instructor could be supported with, for example, data about the trainees' performance in the past, standard explanations for often occurring errors, or system warnings when trainees are not performing at the expected level or show uncommon behaviour …. Not registering and storing performance data also means that trainee performance and the simulator itself cannot be evaluated over a longer period of time" (p. 22). It is remarkable that the evaluated simulators, costing between 0.5 and 35 million Euro, did not include these essential features. A number of more low- and medium-fidelity driving simulators, however, feature automatic feedback and long-term data storage (e.g., Turpin et al., 2007).

To summarize, current focus in simulation-based (driver) training primarily revolves around fidelity and corresponding hardware specifications, whereas the true potential of simulation remains relatively unemployed.

4. Problem statement and aims

Young drivers' road traffic crashes are a major public health concern. The introduction of various engineering and enforcement measures has led to substantial improvement of road safety in the high-income countries. On-road driver training, on the other hand, is generally regarded as safety-ineffective. Different causes of this ineffectiveness have been reported in the literature: higher order skills are insufficiently trained, training is psychologically unsound, there are insufficient opportunities to practise, and there is a lack of standardization and objectivity in assessment and feedback. Driving simulators are increasingly recognized throughout the world as tools for training and assessment. In the Netherlands, about 100 simulators are used for initial pre-licence driver training (Kappé & Van Emmerik, 2005). It is important to recognize advantages of simulators, which include free control over the training conditions, standardization, augmented feedback and instruction, and objective performance measurement.

We recognize the problem that these advantages of driving simulators are not optimally exploited. This thesis will focus on driver assessment in particular. Simulators offer accurate measurements of driver performance; every second, a simulator provides an abundance of information from sensor systems and from the simulated vehicles. It is possible to calculate simple measures related to steering and speed control, speed limit exceedances, turn indicator usage, lane deviations, time-to-collision, and so forth. As explained by Allen et al. (2005b), various algorithms can be applied to these measures, such as means, standard deviations, or power spectra. Based on these performance measures, a Virtual Driving Instructor (also referred to as Intelligent Tutoring System or Personal Information Assistant) can provide automatic feedback and instructions for student-adaptive training. However, the myriad of possible performance measures are intricately related and often redundant. The question remains what performance measures truly convey about the driver. As outlined by Brookhuis (2008), the challenge for the ergonomists in the coming decades is collecting information from operators and the operating environment, and integrating and filtering this information, in order to provide adaptive communication and driver support.

The first aim of this thesis is *to develop a method that can process data obtained from driving simulators into a valid student-profile on driving skill and driving style.* Research is conducted that investigates how to get the most out of objective driver data with the aim of answering the question: Who is the student? What are his or her strengths and weaknesses?

Current research and development in simulation-based training (also including defence and aviation) makes continuing progress in the area of hardware and software, so as to provide a more realistic virtual experience. We recognize the problem that no agreement exists regarding what levels of fidelity (i.e., realism) are required for effective simulation-based driver training. Therefore, the second aim of this thesis is *to develop fundamental knowledge regarding how simulator fidelity relates to training effectiveness*. Motion cueing represents one of the most widely reported discrepancies between simulation and the operational environment and is a major cost driver in the simulation industries. Therefore, this thesis specifically focuses on *investigating the value of low-cost motion cueing systems with respect to training effectiveness*.

5. Thesis outline

The thesis includes ten chapters and two appendixes. Chapters 2–9 have been included with minor changes from their original publications in journals or conference proceedings. This makes the chapters fairly self-contained so that they can be read in random order. Some of the chapters have been updated to make them more consistent with the current state-of-the-art.

The following chapters are included:

2. *Towards a model for deriving measures of individual driver behaviour*
   This chapter provides a literature review of driver behaviour models, with the aim of investigating what type of model/method is most suitable for constructing a student-profile.

3. *Violations and errors during simulation-based driver training*
   This chapter provides a first step toward the construction of a student-profile by using the statistical technique factor analysis. A violation factor, an error factor, and a speed factor are extracted from student performance records.

4. *Relationships between driving simulator performance and driving test results*
   This chapter aims to investigate the relationships between the factor scores derived from simulator performance on the one hand, and the on-road driving test, on the other. Hence, this chapter aims to provide support for the factor model extracted in chapter 3.

5. *Advancing simulation-based driver training: Lessons learned and future perspectives*
   This chapter describes how to improve the effectiveness of simulation-based driver training, evaluates the construction of a student-profile, and describes how the results in this thesis are used in current driver training curricula in the Netherlands.

6. *Driving simulator fidelity and training effectiveness: A literature study of stereo presentations*
   This chapter includes a literature study to obtain insight into the relationships between the realism of a driving simulator and the effectiveness of training, including a case study on the value of stereo presentations. This chapter provides a framework for decision-making with respect to driving simulator fidelity.

7. *The fun of engineering: A motion seat in a driving simulator &*
8. *The search for higher fidelity in fixed-base driving simulation: Six feedback systems evaluated*
   These chapters describe experiments in which we test seven low-cost systems for providing motion information to the driver. It is investigated whether such systems can act as a substitute for more complex motion systems.

9. *Feedback on mirror-checking during simulation-based driver training*
   This chapter includes a study that is independent from the other chapters. As mentioned above, driving simulators do not assess viewing behaviour and consequently cannot provide feedback on important tasks such as mirror-checking. It has been suggested that this is detrimental to training effectiveness (Kappé & Van Emmerik, 2005; SWOV, 2006), but no objective evidence exists for this assertion. For that reason, we performed an experiment that investigates the value of feedback on mirror checking during simulation-based driver training.

10. *Conclusion, discussion, and recommendations*
    This chapter provides conclusions, discusses the results in this thesis, and places these results in context of the theoretical outlines presented in chapters 2 and 6. Future perspectives are presented for improving simulation-based driver training.

The following studies do not explicitly involve driving simulators and have therefore been included as appendixes:

Appendix A. *Gender differences in driving licence theory test scores in the Netherlands*
This study investigates gender differences during the driving licence theory test and discusses the importance of standardization in testing of drivers.

Appendix B. *Exploratory factor analysis with small sample sizes*
The results of chapters 2, 3, and 4 rely to a large extent on the use of the statistical method exploratory factor analysis (EFA). Unfortunately, the available books, manuals of software packages, and many scientific studies do not explain in sufficient detail the conditions under which EFA should yield reliable results. For example, it is often recommended to have a high ratio between the sample size and the number of variables (e.g., Lingard & Rowlinson, 2006), whereas others state that this proposition is false (MacCallum et al., 1999). In addition, literature is inconsistent with respect to sample size requirements. Therefore, a series of simulations were conducted, extending the fundamental knowledge of the conditions under which it is possible to conduct a reliable factor analysis.

CHAPTER 2

Towards a model of driver behaviour

Abstract

Intelligent tutoring systems in driving simulators need to process vast amounts of raw sensor data into reliable and valid indicators about the learner driver in order to provide personalized feedback, instructions, and assistance. The aim of this chapter is to investigate what type of driver behaviour model should be used for constructing a driver profile. Based on a literature survey, and illustrated by four driving simulator experiments, this chapter evaluates the potential of three categories of driver behaviour models: motivational models, adaptive control models, and trait models. It is shown that motivational models, although comprehensive, lack specificity, which makes them inadequate for constructing a driver profile. Adaptive control models, although quantitative and precise, are often overly specific and psychologically implausible. Trait models have been criticized for relying on simple correlations, without incorporating a theory or a multifactorial structural approach. We propose an improved approach to trait models by using the statistical method exploratory factor analysis (EFA). An example is provided of how EFA can be used for driver assessment by extracting a speed factor from a series of diverse performance measures in a complex driving task in a driving simulator. It is concluded that EFA is a useful tool in adaptive support and potentially valuable in further developments towards a generic model of driver behaviour.

De Winter, J.C.F. (2008). *Towards a model of driver behaviour*. Unpublished manuscript.

1. Introduction

Driving simulators are increasingly used as a tool for pre-licence driver training. They allow for complete control over the environment and for objective driver assessment. The data of a simulator can be used for the diagnosis of student competence, statistical analyses, and predictions of future driving performance (e.g., De Winter et al., 2006b). Some driving simulators make use of intelligent tutoring systems to provide feedback and instructions (Kuiken et al., 1992; Weevers et al., 2003b). At the same time, advanced driver assistance systems (ADAS) have been introduced on the market. ADAS provide some similar possibilities for feedback and instruction, but on the road. According to Carsten (2007), the potential of a system that "understands" the driver would be huge. It could give feedback and assistance to the driver, or adapt the operation of the vehicle according to the driver's needs.

A driving simulator or an ADAS can produce vast amounts of raw data regarding an individual's driving behaviour. The challenge that researchers face is how to translate these raw data to valuable and useful measures of the driver and how to construct a student-profile. A valid model of driver behaviour can assist in this process by capturing meaningful patterns and allow a systematic approach to adaptation. During the last decades, a large number of driver models have been proposed in the literature (Carsten, 2007). Modelling driver behaviour has attracted the interest from both the engineering and psychology communities, and the discrepancy of their models has generated interesting debates (Brackstone & McDonald, 1999). Yet, there exists no agreed general framework or a properly validated model with respect to driver behaviour.

The present chapter provides a literature study that investigates what category of model is most useful for generating valid measures of driver behaviour. We adopt the driver model classification by Michon (1985), who proposed a two-dimensional arrangement (behaviourally oriented vs. psychologically oriented, and taxonomic vs. functional) to define four basic types of driver behaviour models: (a) task analyses, (b) trait models, (c) mechanistic/adaptive control models, and (d) motivational/cognitive models. Michon's driver model categories are illustrated in Figure 1.

Michon (1985) defined a taxonomic model as an inventory of facts. The relations in a taxonomic model are those of pertinent structure, for example, subordination, identification, sequential relations, proportions, and likelihood. Contrary to taxonomic models, the components of a functional model interact dynamically; a functional model describes what a driver actually does in traffic situations. Input-output models focus on observable behaviour and mathematical functions. Psychological models, on the other hand, are based on presumptions regarding the processes taking place inside the head of the driver.

	Taxonomic	Functional
Input-output (Behavioural)	Task analyses	Mechanistic models Adaptive control models - Servo-control - Information flow control
Internal state (Psychological)	Trait models	Motivational models Cognitive (process) models

Figure 1. Summary of driver behaviour model types (taken from Michon, 1985).

From Michon, J.A. (1985). A critical view of driver behavior models: What do we know, what should we do? In L. Evans & R.C. Schwing (Eds.), Human behavior and traffic safety (p. 490, Figure 3), New York: Plenum. Reprinted with kind permission of Springer Science and Business Media.

The present chapter evaluates three types of models: (a) motivational models, (b) adaptive control models [1], and (c) trait models [2]. The aim is to investigate what type of driver model could be the most effective for generating valid and useful measures of driver behaviour. The strength and weaknesses of the different categories of models will be evaluated using information available in the literature. In addition, the results of previously conducted experiments in driving simulators will be used as an illustration to the line of reasoning in the text.

2. Motivational models

Motivational driver models are defined as psychological constructs that make explicit assumptions about a driver's internal or mental state (Michon, 1985). Motivational models have qualitative components that dynamically interact. A recent overview of driver behaviour models shows that motivational models of various forms are widely used (Cacciabue, 2007). A thorough chronological overview of motiva-

[1] Michon (1985) used the term "adaptive control models" to characterize models that include functional components that capture behavioural changes. The more traditional definition of adaptive control refers to a class of controllers that have the possibility of modifying their own parameters, for example to facilitate learning (e.g., Flach, 1990). In the present study, we follow Michon's definition and regard an adaptive control model as any functional computational model that makes assumptions about driver behaviour. This also comprises models that only include the (simplistic) driver's aim to minimize speed difference with the car in front.

[2] Task analyses are not evaluated in the present chapter. A task analysis is essentially an inventory of facts about driving tasks, performance objectives, and ability requirements. Task analyses can be used for driver assessment by comparing the driver's performance to the ability requirements of the driving task (i.e., the norms). However, a central thesis of this work (see chapter 1 and experiment 4 in this chapter) is that driver assessment should go beyond such single performance measures in order to recover what lies underneath the data. Therefore, using a trait model in *combination* with a task analysis could be a useful strategy for constructing a driver-profile (Michon, 1985; Quenault et al., 1968; chapters 3 & 4).

tional models was provided by Vaa (2007), starting with the field of safe travel concept of Gibson and Crooks (1938) and ending with the task difficulty model of Fuller (2005).

The risk homeostasis theory (RHT) is likely the most familiar motivational model (Wilde, 1988). The RHT posits that a driver acts based on a target level of accident risk. This target risk level is determined by four subjective utility factors: (a) the expected advantages of risky behaviours, (b) the expected costs of these, (c) the expected benefits of cautious behaviours, and (d) the expected costs of these. At any moment of time, the road user compares his or her personal target level of risk with the level of risk experienced or anticipated and attempts to reduce any difference to zero. The theory further states that whenever a technical, educational, or other intervention is introduced that does not alter the target level of risk, short-term fluctuations in the traffic accident loss per capita may occur, but these will be eventually eliminated after a small delay so that crash rate returns to the pre-intervention level (Wilde, 1988). The RHT has triggered extensive debates that continue in recent overviews of driver models (Cacciabue, 2007; Wilde et al., 2002). Although the RHT has been criticized (e.g., O'Neill & Williams, 2004), it cannot be considered unproductive, as it contributed in bringing the concept of behavioural adaptation and the use of driver incentives on the agenda of the traffic safety research community (Elvik & Vaa, 2004; Trimpop, 1996).

A more recent and well-cited motivational model is Fuller's (2005) task-capability interface (TCI), which describes the interaction between the determinants of task demand and driver capability. Herein, task difficulty homeostasis is proposed as a key subgoal in driving and the choice of speed is argued to be the main solution to the problem of keeping task difficulty within the driver-preferred bounds.

Motivational models often take the form of a comprehensive theory, covering the entire driving task, which could be regarded as evidence for their totality. Opponents, however, have been critical about the overreliance on confirmation instead of refutation, lack of specificity (Ranney, 1994), and impreciseness (Rothengatter, 2002). Lack of specificity can lead to underdetermination of the models, meaning that there exist rival models that are inconsistent with each other, but that are both consistent with the available evidence. For example, the RHT has been found to compete with the zero-risk theory, which argues that subjective risk is *zero* most of the time when driving (Näätänen & Summala, 1974). Janssen and Tenkink (1988) argued that phenomena presented as supportive for the RHT can also be adequately explained in a utility-maximization model. Indeed, multi-utility theories of car driving have been proposed earlier (e.g., Blomquist, 1986; O'Neill, 1977). Another uncertainty has been whether the RHT applies to individual drivers or to aggregate behaviour of a population of drivers (Huguenin & Rumar, 2001). Clearly, it is important to accurately distinguish between individual and collective behaviours. For example, in the quantitative field of elementary reaction time tasks, the speed-accuracy relation is well known as a within-subjects trade-off. However, between subjects, a positive correlation has been found. That is, the persons who react more quickly are also generally those who make fewer errors (Jensen, 2006). The critiques about the lack of specificity do not apply only to models that use risk as the primary mental state but to other motivational constructs as well. For example, the popular concept of *affordances*, which can be used to characterize Gibson's field of safe travel (Gibson & Crooks, 1938; Vicente, 1999), has received comparable account, namely that the concept is too ambiguous (Oliver, 2005).

As a response to the risk-oriented models, Ranney (1994) proposed hierarchical models, which he also referred to as the second generation of motivational models. Literature contains several driver hierarchies, mostly comprising three levels (e.g., Michon, 1985; Rasmussen, 1983; Van der Molen & Bötticher, 1988) or four levels (e.g., Hatakka et al., 2002; Hollnagel et al., 2003; Panou et al., 2007). Researchers have also proposed to combine hierarchies or taxonomies into a two-dimensional matrix (Hale et al., 1990) or a three-dimensional cube (Summala, 1996; Theeuwes, 2001). However, lack of specificity remains a concern. Hierarchies are often presented without thorough quantitative support. Moreover, it remains unclear how the multiple-level hierarchies should be interpreted in relation to two-level distinctions that are also found in literature. Examples are accuracy and speed (Zhai et al., 2004), errors and violations (Reason et al., 1990), driving skills and driving style (Elander et al., 1993), driver performance and driver behaviour (Evans, 2004), skills and safety-motives, (Lajunen & Summala, 1995), and automatic and willed control of behaviour (Norman & Shallice, 1986).

Experiment 1. The RHT and TCI during simulation-based training

As mentioned above, motivational models are attractive because of their entirety. Suppose that the RHT would be valid during driving. An individual's target risk level would then most likely be *the* all-embracing variable determining driver behaviour and therefore a very valuable indicator to be used in driver-adaptive feedback and instructions.

We reanalysed data of a previous study (De Winter et al., 2006d) to investigate whether the RHT or the TCI are valid during simulation-based driver training. The experiment was conducted in a fixed-base driving simulator (Green Dino, 2007). Ten participants (8 men, 2 women; age range 18–30 years) without previous driving experience were selected. Each participant completed four identical sessions. In each session, they drove a 7.6 km lap on a two-lane rural road with 25 bends of varying radii. Lane width was 5 m. Participants had to steer, apply throttle, and brake by themselves; gear changing was automated. Participants were instructed to drive as well as they could to keep in the right lane with both hands on the steering wheel, and to adopt a speed of 20–30 km/h in sharp corners. During the simulated drive, no other vehicles were present, and no feedback was provided. Thus,

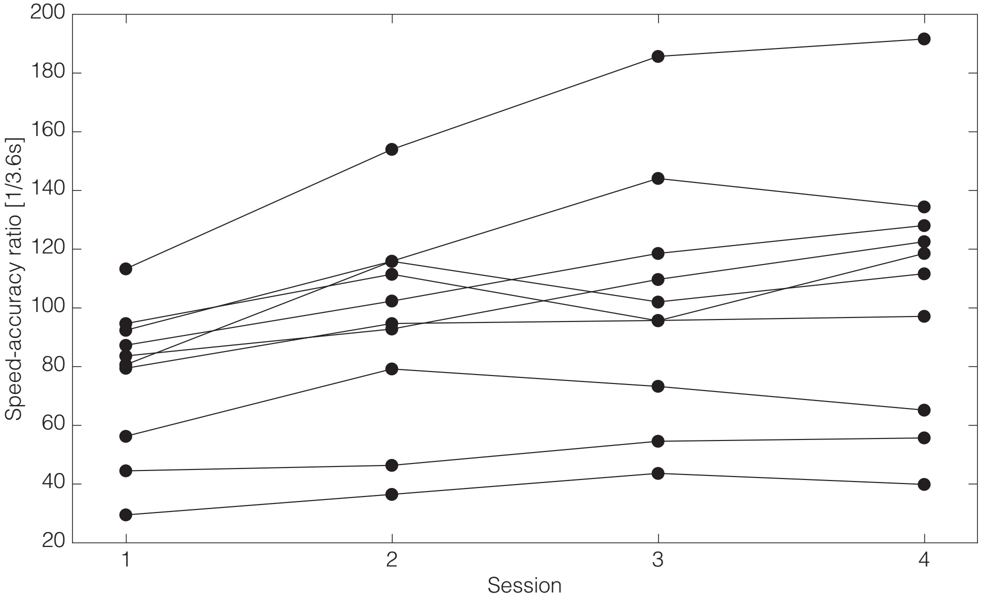


Figure 2. Speed-accuracy ratio (mean speed [km/h] divided by the standard deviation of lateral position [m]) of the 10 participants as a function of session number.

the experiment was a simple, self-paced training of the elementary task of driving the car around a track.

Objective measures included the mean speed and the lane keeping accuracy measured by the standard deviation of lateral position, or SDLP for short. These measures were calculated for each road segment separately and subsequently averaged. After each session, participants completed a written questionnaire with a series of 16-point scales assessing their experience in the preceding session. These selected questions under the present investigation were (translated from Dutch): 1) "How difficult did you find the task during the preceding session?", from 1 *(very easy*) to 16 (*very difficult*). This question had to be answered with respect to straights, normal turns, sharp turns, and mild turns separately, and was averaged into one task difficulty level, and 2) "According to your own feeling, how high did you regard your risk level during the preceding session?", from 1 (*high*) to 16 (*low*). Afterwards, all measures were transformed to a scale running from 0% (*very easy/low risk*) to 100% (*very difficult/high risk*).

The results of a previous analysis of these data (De Winter et al., 2006d) showed that the participants improved performance with session number. Depending on the participant, this improvement was expressed in two separate ways: increasing speed or improving accuracy. Figure 2 shows the speed-accuracy ratio (mean speed divided by SDLP) of the 10 participants as a function of session number. Participants improved on this aggregate performance indicator (*t* = -4.52, *p* = 0.001, comparing session 1 and 4 using a paired *t* test).

Figure 3 shows the results of the subjective measures. Task difficulty decreased (*t* = -6.18, *p* < 0.001) as well as participants' feeling of risk (*t* = -2.93, *p* = 0.02). Hence, the questionnaire results showed that the task became subjectively easier and less risky with increasing driving experience.

The results of the experiment showed that subjective risk and subjective task difficulty decreased with increase of experience. These results seem in strong disagreement with the homeostatic core of the RHT and the TCI. However, the present experiment *does not* falsify either the RHT or the TCI because these models do not explicitly state that risk or task difficulty should always be constant. The TCI includes a comparator between perceived task difficulty and the range of acceptable task difficulties in order to keep the perceived task difficulty within this range. The model does not specify how wide this range can be. One may argue that the experience increased driver competence, which in turn reduced task difficulty. Suppose, however, that constant task difficulty was observed, then one could claim that the TCI predicted that the driver kept task difficulty constant by behavioural compensation (e.g., by reducing effort or increasing speed). In other words, the problem is

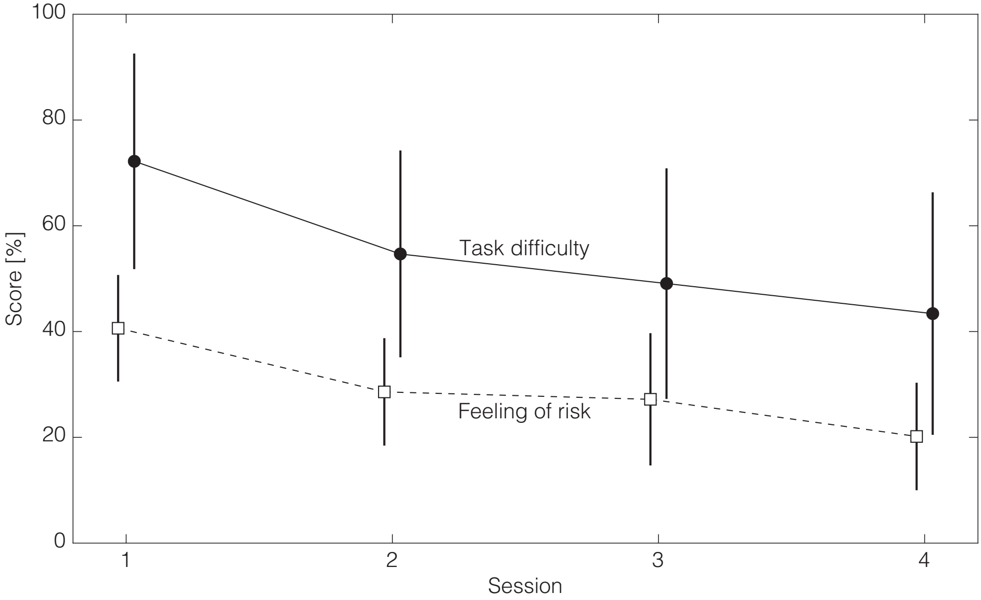


Figure 3. Mean questionnaire responses as a function of session number. The error bars are depicted at ±1 standard deviation from the mean. The results have been offset slightly on the horizontal axis so that overlap of the error bars can be seen more clearly.

that it is *always* possible to come up with a reason why the risk level or task difficulty level changed and by this means save the models from falsification (see Elvik & Vaa, 2004; Huguenin, 1988; Ranney, 1994, for similar comments regarding the RHT and other motivational models). To summarize, the present section described the problem of unfalsifiability: because the motivational models are all-encompassing without accurate specifications of boundary conditions, it is difficult to use them for making testable predictions, or to apply them for driver-profiling.

3. Adaptive control models

Another branch of driver models is formed by adaptive control models, which, according to Michon (1985), can be subdivided into classical control models, developed in the context of the processing of signals that are continuous in time, and information-flow control models, involving discrete decisions. Adaptive control models range from microscopic engineering models (Bracktone & McDonald, 1999), expressions that describe the human as a manual controller (e.g., McRuer & Jex, 1967), to intricate simulation models (see e.g., Cacciabue, 2007). Adaptive control models have been extensively used for the identification of driver parameters, such as a driver's control characteristics and time delays, based on measured performance data (e.g., Kano et al., 2005). Overviews of various types of adaptive control models are provided by Brackstone and McDonald (1999), Guo and Guan (1993), Jagacinsky and Flach (2003), MacAdam (2003), Plöchl and Edelmann (2007), and Reid (1983). Opposite to motivational models, adaptive control models provide specific and accurate results. A purported drawback, however, is that they have been successfully applied almost exclusively to isolated driving tasks, such as taking curves, single-lane car-following, lane changing, and obstacle avoidance.

Adaptive control models are mostly applied by engineers, who tend to adopt a different approach to driver modelling than psychologists. This became strikingly apparent after a publication of an overview on engineering-oriented car-following models by Brackstone and McDonald (1999) in a psychologically-oriented journal. Although it became a much-cited article, there was considerable scepticism as well. The review generated four commentaries (Boer, 1999; Hancock, 1999; Ranney, 1999; Van Winsum, 1999). Ranney (1999) criticized that Brackstone and McDonald made very sophisticated assumptions that were not well-motivated from a human perspective. Many factors that are known to influence car-following, such as weather, road conditions, age, gender, motivations for driving, strategic aspects, and the notion that drivers may be satisfied with a range of conditions and do not behave optimally, were not covered. Hancock (1999) criticized the fit-driven approach of the models: "Deriving equations from physical descriptions of motion and subsequently trying to fit these to data derived from behavioral response both literally and figuratively, puts the cart before the horse." (p. 198). Boer (1999) considered that the aspect of individual differences was overlooked: "Psychologically mediated situation dependent within and between driver variability is a likely cause for the lack of agreement between the multitude of car-following models" (p. 205) and that "most driver models assume that drivers have access to variables that are mathematically convenient but perceptually implausible" (Boer et al., 2000). To summarize, the models in the engineering-oriented review of Brackstone and McDonald (1999) were considered questionable from a psychological point of view.

Experiment 2. Redundant measures

Adaptive control models often act on an error or disturbance, such as speed and distance with respect to the vehicle in front (in a car-following task), or distance and heading error with respect to the desired vehicle path (in a lane keeping task). Therefore, adaptive control models cannot make individual predictions in the absence of disturbances from the environment. However, even in stationary conditions, within- and between-subjects variability can be significant.

To illustrate this, data were reanalysed of a previously conducted car-following experiment in a driving simulator with 10 licensed drivers (2 women, 8 men; age range 18–34 years; Kouwenberg, 2005). The initial aim of the experiment was to investigate driver attention and performance during a 1 hour car-following task. Participants were instructed to follow a lead vehicle at a constant comfortable distance. The lead vehicle had a constant speed of 100 km/h. No other traffic was present. The experiment took place on an endless straight road with constant lead vehicle speed; the virtual situation can therefore be considered entirely stationary. Background music was played during the experiment. Half of the experiment involved a secondary task (rhythmic tapping), but this had only small effects on driver performance.

Figure 4 shows that there were considerable individual differences in following distance and distance variability. The following distance within participants varied as well, as indicated by the standard deviation of following distance being larger than 10 m for each participant. In other words, even in stationary conditions there was significant inter- and intrasubject variability in driving behaviour.

A positive correlation between the mean and standard deviation of following distance can be inferred from Figure 4 ($r = 0.89$, $p < 0.001$). This correlation demonstrates that performance measures can be redundant. Previous studies have found a similar positive correlation between mean and standard deviation of intervehicle spacing distance during (nonstationary) car-following and have used adaptive control models to help explaining this phenomenon (e.g., Boer et al., 2005; Mulder et al., 2005). However, different models have been used, and different explanations for this correlation have been described, which illustrates that even a single interindividual correlation is not straightforwardly depicted by an adaptive control model.

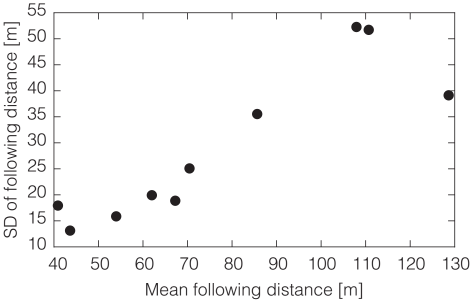


Figure 4. Standard deviation of following distance versus mean following distance of 10 participants who completed a car-following experiment.

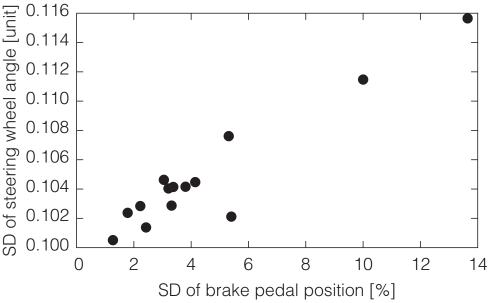


Figure 5. Standard deviation of steering wheel angle versus standard deviation of brake pedal position of 14 participants who completed a car-following experiment.

Experiment 3. Psychologically mediated relationships

The process of connecting information and functions in an adaptive control model is inevitably somewhat arbitrary. The way that various transfer functions are connected depends on the researcher's own choices. It is always possible to construct a model and to subsequently fit the model to the data. However, this does not mean that the model provides a good representation of driver behaviour. Many adaptive control models adopt a distinction between longitudinal and lateral driver behaviour (e.g., MacAdam, 2003). From the standpoint of modelling the physical vehicular response dynamics, such a distinction is often indeed justified. However, when trying to describe individual differences, other models may be more appropriate.

To illustrate this, data were acquired from a previous car-following experiment in a driving simulator (De Winter et al., 2006c, 2008f). Fourteen participants (12 men, 2 women; age range 23–51 years; all having car driving experience) followed a preceding car in traffic in four 12-minute sessions on a highway, while testing various configurations of an accelerator force feedback system. At distinct moments, vehicles cut in from the left lane, just in front of the participant's car. Results showed that participants who were more active with their brake pedal were more active with their steering wheel as well ($r = 0.95$, $p < 0.001$) (see Figure 5). In a statistical factor analysis, these behaviours were grouped under the same factor (De Winter et al., 2006c), leading us to hypothesize that both forms of physical activity were governed by the same latent psychological phenomenon. Although the sample size in this experiment was rather small, this study clearly showed that a model structure, such as a distinction between longitudinal and lateral behaviour, should not be taken for granted. Latent unobserved psychological factors may be present as well, coupling the previously uncoupled dynamics.

In the same spirit, Boer (1999) made some critical notes with respect to the generalizability of the engineering models of the Brackstone and McDonald review (1999): “Experience tells us that the adopted modeling abstraction has been pushed too far as exemplified by the fact that the identified coefficients of the same model differ considerably between experiments” (p. 202). Indeed, a common characteristic of adaptive control models is that they are often well able to fit accurately measured data. Lack of validation remains problematic, however.

Many studies expressed recommendations to further *increase* the level of complexity of the models. McRuer et al. (1977) discussed extensions and refinement, such as including model components for driver decisions and driver judgement factors and interactions. Guo and Guan (1993) recommended further improvements, such as better preview and correction strategies in order to render better tracking accuracy. Brackstone and McDonald (1999) discussed extending the car-following models to include motivational and attitudinal factors. MacAdam (2003) concluded that more refined and accurate models should include more functionality, such as driver sensory input processing elements. According to Michon (1985), a model that displays a sufficiently broad spectrum of realistic driving behaviours will be inherently complex and will embody at least between 5,000 and 10,000 elements.

Adequate fit to observed data is a necessary but not a sufficient condition for model quality (Pitt et al., 2002). In essence, a model should carry predictive validity. That is, researchers are interested not only in how well a model can describe a process but also in how well the model with its parameters constrained to fixed values can describe other (e.g., future) data. Models of higher complexity (i.e., more parameters, functions, connections) are generally more flexible to describe the data but at a price of reduced predictive validity. Essentially, the risk of an overfitted model is that its driver parameters become meaningless. For example, the Optimal Control Model of human behaviour (Kleinman et al., 1970) contains many free parameters and it has been concluded that this considerably hampers its use and predictive ability (Mulder, 1999). Good discussions on overfitting in scientific models can be found in Cutting (2000), Ginzburg and Jensen (2004), and Preacher (2003). Indeed, the success of the simple two-parameter Crossover Model of McRuer (McRuer & Jex, 1967), for example, has been largely attributed to its simplicity (De Winter et al., 2008e). Although simple and plausible adaptive control models exist (e.g., Allen et al., 2005a; Boer et al., 2005), they have been applied only in closely controlled elementary driving tasks; not in varied realistic driving tasks.

4. Trait models

Trait models describe relations or categorizations regarding individual differences in driving behaviour, without incorporating dynamic functional components. Trait models have generally evolved around the statistical identification of accident-prone drivers. Michon (1985) was critical towards trait models and stated that empirical connections are at best correlative. Ranney (1994) provided a review about the study of individual differences in road traffic crashes and concluded that the differential crash involvement paradigm should be abandoned.

Ranney (1994) was particularly critical about post hoc explanations and the absence of a multifactorial structural approach when evaluating relations between measurements. Quoting Ranney (1994): "The use of simple correlational methods without multifactorial structural models raises questions about the meaning of significant correlations" (p. 736). Present literature on driver assessment employs many different performance measures (e.g., Östlund et al., 2004). The results above showed that driver measures can be redundant. Some of these correlations may be explained by a careful study of the vehicle dynamic response, whereas others may be mediated through psychological influences. Although many performance measures have been successfully used, the question that again arises is: what do they really convey about the driver?

Another criticism of Ranney (1994) was that crashes are problematic as a criterion. Car crashes are infrequent events, and the statistical power is generally very low. Furthermore, it is difficult to reliably and consistently measure car crashes for a group because not all crashes are observed or reported, and many external factors play a role, such as regional climate differences and law enforcement practices. However, the fact that a criterion is unreliable, or that something cannot be easily measured, is no reason to devalue the trait-based modelling approach as such. Rather, alternative driver safety measures have to be sought.

Considering the inherent limitations of motivational and adaptive control models and despite the lack of success of trait models in the past, statistical models of driver behaviour have continued to receive attention. Elvik and Vaa (2004) deduced that explaining driver crashes is inevitably a statistical procedure. In a recent overview about driver models, Carsten (2007) stated that one should not use a deterministic approach for driver modelling and proceeded to identify the psychological factors governing driving behaviour. Rothengatter (2002) proposed that the study of individual differences can provide a basis for driver training and accident prevention.

There is nothing fundamentally wrong with trait-based driver behaviour modelling, as long as one applies the right statistical techniques. We propose to use exploratory factor analysis (EFA) to investigate what common factors can explain correlations between various measures of a driver behaviour. EFA is a method that does not impose causal relations, but instead it *uses* the correlations amongst measures to identify a smaller number of underlying factors that together parsimoniously explain the data.

Table 1. Performance measures calculated for each participant

TTImin [s]	This safety indicator represents the participant's time-to-intersection (TTI) at the latest moment that both vehicles were on a collision course. A collision course was defined as a difference between the TTI's of both drivers of less than 2 s. This measure provided an indication of the intervehicle conflict. A lower TTImin represents a higher intervehicle conflict.
PET [s]	The post encroachment time (PET) represents the absolute time difference between two interaction partners reaching and leaving the point where their paths intersected.
TTIdiff [–]	The base-10 logarithm of the difference between the TTI's of both drivers at the moment when the participant is 15 m from the centre of the intersection. The smaller the TTIdiff score, the more dangerous the interaction (adapted from De Winter et al., 2006a).
MinDist [m]	The minimum distance between both vehicles
MeanSpeed [km/h]	Mean speed
MinSpeed [km/h]	Minimum speed
Yield [0 or 1]	Indicates whether the participant gave right of way to the other driver (1) or not (0)
BrakeTime [s]	Total time that the brake was pressed
NoThrottleTime [s]	Total time that the throttle was released
DTIini [m]	Distance to the intersection at onset of braking
Vini [km/h]	Speed of the vehicle at onset of braking
BrakeMax [0–1]	Maximum brake pedal position
ThrottleMax [0–1]	Maximum throttle position

Note. Each measure was averaged over all intersections and all sessions completed by the participant.

Experiment 4. Applying factor analysis in a complex driving task

In order to illustrate the possible merit of EFA in driver modelling, we obtained performance data on 25 drivers who participated in a simulator experiment. The initial goal of the experiment was to study interactive driving behaviour at intersections and different ways, in particular the provision of visual-auditory information, to influence this behaviour (Houtenbos, 2008). All participants (17 men, 8 women; age range 25–70) had their driving licence for at least five years and had driven at least 5,000 km in the previous year.

Participants completed four sessions in a virtual environment with 20 intersections. Participants started with two short practice trials and continued with a session with information turned off, followed by two information sessions, and completed the experiment with a session with information turned off again. In order to facilitate natural interactive behaviour between vehicles on the intersections, the simulator

Table 2. Descriptive statistics and factor loadings of participants (*N* = 25)

	Mean	SD	Min	Max	Loading
TTImin [s]	3.70	1.29	2.00	7.82	-0.96
PET [s]	2.48	0.49	1.79	4.27	-0.83
TTIdiff [–]	0.91	0.39	0.42	2.13	-0.50
MinDist [m]	23.2	3.19	18.6	32.5	-0.82
MeanSpeed [km/h]	45.0	3.52	34.1	52.0	0.82
MinSpeed [km/h]	31.4	5.76	14.4	40.4	0.96
Yield [0 or 1]	0.41	0.05	0.29	0.50	-0.39
BrakeTime [s]	1.19	0.55	0.42	3.20	-0.85
NoThrottleTime [s]	5.15	1.52	2.21	9.54	-0.86
DTIini [m]	53.3	9.03	34.4	79.7	-0.75
Vini [km/h]	47.3	3.65	43.8	59.8	0.32
BrakeMax [0–1]	0.23	0.07	0.11	0.41	-0.41
ThrottleMax [0–1]	0.48	0.14	0.28	0.83	-0.26

was coupled with a second simulator, driven by a human experimenter. That is, two drivers encountered each other in the same virtual world. Participants were instructed to keep the speed limit of 50 km/h. At intersections, the experimenter's vehicle could either (a) not appear, (b) appear from the left while maintaining speed, (c) appear from the left while slowing down, (d) appear from the right while maintaining speed, and (e) appear from the right while slowing down. Hence, the experiment provided data on 25 individuals who all had negotiated 80 intersections. Driver behaviour was complex and dependent on the type of situation, individual differences, visibility (intervehicle visibility was manipulated by buildings placed in the line of sight), interactive behaviour between two human drivers, and possibly also strategic decisions whether or not to take right of way and how to preserve an energy-efficient drive.

Thirteen selected performance measures of intersection behaviour were calculated, shown in Table 1. These were safety-measures, speed-related measures, and measures describing at which position the participant released the throttle, how long he or she pressed the brake, and so forth. Table 2 provides descriptive statistics for each measure. Next, a correlation matrix was constructed amongst the 13 measures and submitted to principal-axis EFA. Based on the scree plot we decided to extract one factor. The first eigenvalue was 6.95; the second and third, unretained eigenvalues were 1.99 and 1.35, respectively. Note that these circumstances (relatively small sample size but high communalities and one extracted factor) yield reliable EFA results (appendix B).

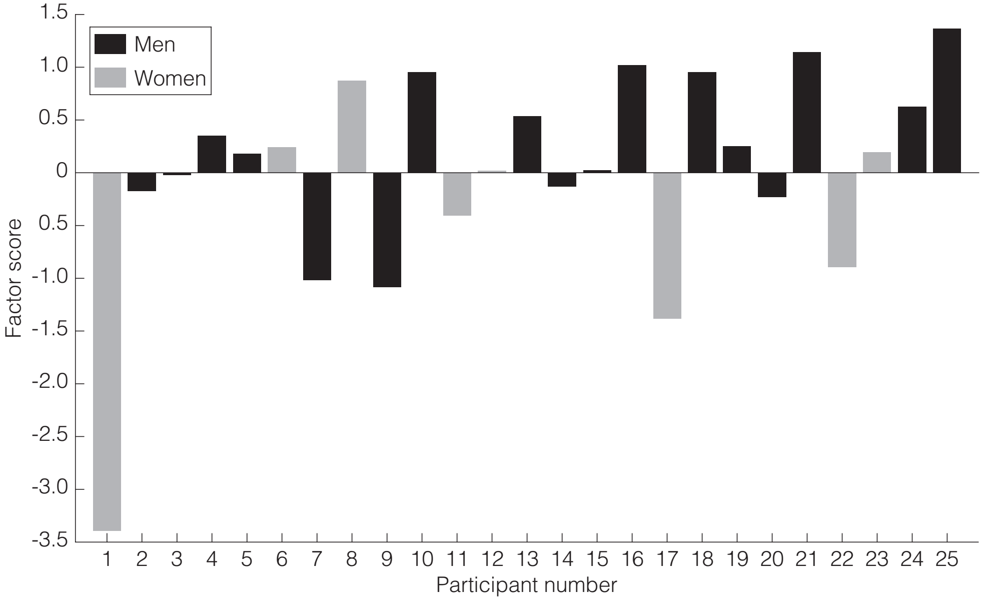


Figure 6. Factor scores for the 25 participants.

Table 2 shows the loadings of the one factor solution; this solution explained 51% of the total variance. That is, a substantial share of the variability of the manifest variables was explained by one common factor; the remainder is the variables' uniqueness or randomness. High factor loadings (i.e., correlation coefficients with the factor) were found for the minimum speed and mean speed, whereas low loadings were found for the safety measures, yielding to the other road user, BrakeTime, NoThrottleTime, DTIini, and BrakeMax. Hence, the extracted factor was interpreted as a general speed factor. The identification of a speed-factor is compatible with Groeger (2000b) and Rothengatter (1988) who considered that speed is a personal characteristic that shows consistency over time and over locations.

Participants' speed-scores were calculated (Figure 6). The speed-score correlated with gender (r = -0.41, p = 0.041 [0 = man, 1 = woman]), age (r = -0.38, p = 0.058), and with the interpersonal violation-score that was calculated from the participants' prior responses to the Driver Behaviour Questionnaire (r = 0.47, p = 0.018; see Houtenbos, 2008 (pp. 177–178); factor structure taken from Mesken et al., 2002). Next, in order to evaluate split-half reliability, the speed-scores were calculated twice: once based on the data of sessions 1 & 3 and once based on the data of sessions 2 & 4. These speed-scores showed a strong correlation (r = 0.96, p < 0.001) indicating adequate reliability.

In the studies described in chapters 3 and 4, a speed-score was extracted from the mean task completion times of diverse driving tasks in a simulator. An important finding was that the speed-score was more reliable and discriminated more effectively between men and women than any of the single measures in isolation. Chapter 4 assesses the predictive validity of the factor-scores by determining the statistical relationships with results of the on-road driving test.

It is concluded that by using EFA, a meaningful and interpretable driver measure could be derived. Such a measure could potentially be applied for constructing a driver profile and for summative or formative assessments in order to improve driver training effectiveness. For instance, a speed-score can be presented by the driving simulator and used for feedback on performance.

5. Discussion

5.1. Motivational, adaptive control, and trait models

Vehicle's sensor systems or computers in driving simulators can produce large quantities of data. In order to provide driver-adaptive feedback and instructions, these data must be processed in a valid way to obtain meaningful measures about the driver. A driver behaviour model can be helpful to assist in this process.

Motivational models have shown their value in generating discussions and ideas regarding the mechanisms that take place in a driver's psyche. However, they have received extensive critiques as well, with lack of specificity being a major concern. In an experiment, we used written questionnaires to quantify participants' subjective task difficulty and feeling of risk during initial simulation-based driver training. Using these results, we discussed the problem that the motivational models are not falsifiable, which makes it difficult to establish trust in their usefulness for creating valid driver measures.

Adaptive control models, on the other hand, have been successfully applied in vehicle dynamics research as well as in driver identification studies to generate precise results. However, they have been mostly applied for specific tasks, with sometimes disconcerting predictive validity and plausibility. We showed that, even under stationary conditions, there were considerable individual differences and difficult-to-model correlations between various performance measures. Moreover, there was a risk of model overfitting; a "dangerous" practice because if you overfit "you think you know more than you really know" (Grünwald, 2000, p. 148). We conclude that it is difficult to employ adaptive control models for generating driver measures that have sufficient generalizability. The usefulness of adaptive control models is restricted to elementary driving tasks paradigms.

Trait models have been criticized in the past for using post hoc explanations without a multifactorial structural approach. We asserted that trait models are promising, but that the right statistical techniques should be used in order to go beyond establishing correlations between single variables.

5.2. The use of factor analysis for extracting driver measures

The statistical technique exploratory factor analysis (EFA) was proposed as an alternative to motivational models, adaptive control models, and the plain correlational trait models. Essentially, EFA is a trait model as well, but it goes beyond being merely correlational, by detecting latent structures in data. An example showed how EFA can be used on intricately related measures, and how it can extract a meaningful measure about driver behaviour, in this case a speed-score. Although the example was limited to intersection crossing behaviour, some important benefits of EFA were demonstrated. It was shown that EFA is able to detect regularity amongst performance measures. We have previously tried to model the same intersection task using adaptive control models of various kinds (e.g., in terms of competitive and cooperative behaviour of road users), but the task was too complex and too varied to yield meaningful results. In addition, the example showed that various measures (e.g., minimum speed, safety indicators) had high loadings on the same speed factor, indicating that they are largely redundant with each other. Factor scores can potentially be applied for feedback to a learner driver or to an ADAS user in order to improve training effectiveness and road safety.

Opposite to adaptive control models, EFA does not impose causality. Instead, it *uses* the correlation matrix to identify a smaller number of underlying factors that explain the data. In this way, EFA exploits an important principle of scientific inference known as the principle of the common cause (Haig, 2005). EFA is not a deductive technique; it can be regarded as an abductive method of science, that is, it derives a logical, parsimonious explanation of data. It has a somewhat heuristic nature, but it is not as vague as motivational models. EFA bears close resemblance to principal component analysis (PCA). According to many, EFA is a more powerful technique than PCA because it has the capacity to better retrieve the constructs that underlie the data. However, in practice, the differences of EFA and PCA are not well understood and are an issue of debate (Steiger, 1994). In most circumstances, EFA and PCA yield near-identical results (Velicer & Jackson, 1990). PCA is better known as a data-reduction technique. That is, contrary to the risk of limited generalizability of complex models, EFA and PCA aid in the pursuit towards simplicity by reducing large amounts of data into fewer factors or components.

Before conducting EFA, several important methodological decisions have to be made. As with any statistical technique, one has to decide first which samples and variables to use. Next, one has to choose which model fit procedure to use, how many factors to extract, which rotation to apply (if any), and which method to use for calculating scores. Each of these decisions has consequences to a varying degree of importance on the results obtained (Fabrigar et al., 1999; Velicer & Jackson, 1990). After conducting EFA, the factors have to be interpreted for their meaning. The results of EFA are not uniquely defined by the data and factorial invariance for different populations or different moments in time are another aspect to be considered. Some theorists have criticized the indeterminacy associated with factor analysis (e.g., Schönemann, 1990). Defenders of factor analysis have responded that indeterminacy of factors is just a special case of the general indeterminacy of theory by empirical evidence widely encountered in science and does not pose a particular problem (Haig, 2005; McDonald & Mulaik, 1979). In this respect, analogies can be made with discussions about models of human intelligence or personality. In psychometric intelligence research, for example, there is considerable disagreement about how many factors underlie test data and whether it is one general intelligence factor or a set of partially independent intelligence factors (Gottfredson, 1997; Neisser et al., 1996). Most likely there exists no single answer as all models fit data to different degrees (Bentler, 2000). A similar notion applies to driving. It is to be expected that a single model of driver behaviour does not exist. However, models can be compared with respect to their parsimony and their predictive value.

The present example was not the first time that EFA has been used in the context of modelling driver behaviour. EFA and PCA have been applied on DBQ responses, resulting in violation and error components (e.g., Reason et al., 1990). Variants of the DBQ have been widely employed in numerous other studies (Åberg & Rimmö, 1998; Blockey & Hartley, 1995; Lajunen et al., 2004). EFA has also been applied for studying phenomena such as decision-making style and driving style (e.g., French et al., 1993), driver aggression and anger (Hauber, 1980; Iversen & Rundmo 2002, 2004), driver stress (Gulian et al., 1989), driver fatigue (Matthews & Desmond, 1997), perceptual motor skills and safety skills (Lajunen et al., 1998), skills and safety-motives (Lajunen & Summala, 1995), problem driving (Hartos et al., 2000), driver vengeance (Wiesenthal et al., 2000), driver emotion (Mesken, 2006), avoidance of driving (Stewart & St. Peter, 2004), and attitudes and habits towards speeding (De Pelsmacker & Janssens, 2005). The majority of EFA studies appear to be primarily based on self-reports of driving without incorporating objective driving performance. A smaller number of studies have applied EFA on objective driving data, however. Some examples are Janssen (1994), Lundqvist et al. (2000), De Winter et al. (2006a) and chapters 3 and 4. A generic model of driver behaviour can integrate questionnaire responses with objective measures.

5.3. Psychological and engineering models

Tentatively, one may conclude that EFA takes a distinct position between unspecific motivational models and overly specific adaptive control models. The car itself is a fairly deterministic system, many aspects of which can be well described and simulated using mathematics, elementary mechanics, and dynamics. Humans are less deterministic, comprising interindividual differences (e.g., skill, age, gender), intraindividual differences (e.g., fatigue, concentration, emotion), physical and mental processes and limitations, learning, and various forms of error. The caveat often seen in this respect is that vague and unfalsifiable models of car driving are presented. As quoted from Hancock (1999): "The larger question remains. Will the physicist and the psychologist ever meet? Someday they must. Yet clearly much remains to be achieved if those of a mathematical persuasion are to change their fundamental perspective to a psychological focus while many in psychology learn to use the austere scalpel of numbers in their descriptions of behavior. Hopefully, such a union will be of great value in transportation and many worlds beyond." (p. 199). EFA is a quantitative but nondeterministic technique and potentially valuable for explaining human car driving, which is irreducibly a *human-machine* process.

5.4. Limitations and other types of driver models

This chapter was inevitably not comprehensive in describing all driver behaviour models that currently exist. Excellent overviews of driver models can be found in Cacciabue (2007), Jagacinski and Flach (2003), Michon (1985), and Ranney (1994). Referring to the taxonomy of Michon (Figure 1), task analyses were not evaluated in the present chapter. Mechanistic models were neither. A mechanistic model is, for example, a model that regards a traffic stream mathematically as a continuous incompressible fluid. Although mechanistic and macroscopic models of traffic behaviour are still widely employed, we agree with Michon's suggestions that drivers most likely do not adhere to such assumptions.

A remark is made with respect to cognitive process models (Michon, 1985). We consider that cognitive models belong better to the category of adaptive control models. Indeed, Michon (1985) refers to cognitive models as adaptive control models of thought. Although cognitive models are of a more sophisticated kind than "mindless" information-flow control models, there was no obvious reason to cluster cognitive models in the same category as motivational models such as the RHT. In fact, concerns about limited generalizability apply to cognitive models as well: It has been pointed out that cognitive theory is radically underdetermined by data (Newell, 1992).

Finally, the present review particularly addressed individual differences in car driving behaviour. The study of within-subjects variability, particularly the use of psychophysiological indicators, is certainly of relevance as well (Brookhuis & De Waard, 1993; Brookhuis et al., 2003). As indicated by Brookhuis (2008): "One of the major problems of an adequate adaptive vehicle control system is to detect and assess inadequate driving by the driver of the motor vehicle; when and why performance drops 'below the red line', where and what exactly is this red line" (p. 58). Answering this question requires a careful study of the dynamics of inter- and intraindividual differences. It is conceivable that EFA can be useful in this respect as well. Physiological measurements, such as eye-tracking data, can be combined with objective measures using (dynamic) EFA to identify their communality.

5.5. A generic model of driver behaviour

A future generic model of driver behaviour can integrate subjective with objective measures. Potential constructs to include in a generic multivariate model are speed, errors, crashes, personality, age, experience, gender, intelligence, sensation seeking, and neurological factors (e.g., Spiers & Maguire, 2007). Following EFA, related multivariate statistical techniques that are directed towards hypothesis testing, such as confirmatory factor analysis (CFA) and structural equation modelling (SEM), can facilitate in constructing and testing a generic driver model. Several important steps have already been undertaken. For instance, Groeger (2000a) tested a four-factor cognitive theory of driving behaviour using CFA. A similar type of study was presented in Grayson et al. (2003). SEM has been applied to driver behaviour by, for example, De Pelsmacker and Janssens (2005), Sato and Akamatsu (2008), Ulleberg and Rundmo (2003), and Wallén Warner and Åberg (2006). The merit of SEM for travel behaviour research and driver behaviour studies has been acknowledged by Golob (2003) as well. Such approaches are constructive in the maturing field of driver behaviour modelling.

6. Conclusion

This chapter investigated the merit of motivational, adaptive control, and trait models. It was concluded that motivational models and adaptive control models were not successful as tools for constructing a driver profile. We proposed to apply the trait models concept by using EFA as a statistical multivariate tool in the endeavour of extracting individual measures of driving behaviour and potentially uniting engineers' and psychologists' perspectives on the world. A generic speed factor could certainly be of importance in a model of driver behaviour, considering that speed plays a central role in many existing driver models.

CHAPTER 3

Violations and errors during simulation-based driver training

Abstract

The effectiveness of virtual driving instruction can increase when techniques that automatically distinguish between violations and errors are available, two behaviours requiring different types of remediation. This study reports the analysis of the objectively measured performance of 520 participants completing a simulation-based training programme. Factor analysis of failure reasons showed that violations and errors were the primary underlying factors. Men committed more violations and women made more errors; the magnitude of gender differences corresponded to the factor loadings. Factor analysis of the mean task completion times yielded a factor that can be described as the extent to which motivation for speed was expressed in quicker task execution. Quicker participants completed more tasks, committed more violations but made fewer errors. Participants reduced errors during forced-paced driving and increased speed during self-paced driving. The authors recommend exploiting the distinction between violations and errors by developing interfaces and feedback for both types of aberration.

De Winter, J.C.F., Wieringa, P.A., Kuipers, J., Mulder, J.A., & Mulder, M. (2007c). Violations and errors during simulation-based driver training. *Ergonomics*. *50*, 138–158. (adapted with minor textual changes)

1. Introduction

In these days of high-priced fuel, simulators provide a cost-effective solution to initial driver training. Besides financial benefits, simulators offer great opportunities for carrying out objective measurements on the user's actions in a safe and purpose-developed virtual environment (Vlakveld, 2005b). Not surprisingly, many driving schools have started using simulators for training their students. It has been estimated that around 100 low-cost simulators are currently used in the Netherlands for initial driver training (Kappé & Van Emmerik, 2005). Nowadays, the introduction of virtual instructor software, sometimes referred to as an intelligent tutoring system, contributes to further automation of driver education, the result of which changes the role of the human instructor into that of a supervisor (Kappé & Van Emmerik, 2005; Weevers et al., 2003a, 2003b).

In comparison with a human instructor, a virtual instructor monitors and assesses the driver's actions by comparing his or her performance with normative performance and gives corrective feedback when the driver fails to comply with these norms (e.g., Michon, 1993). However, there are indications that failures cannot always be effectively remedied by practice and simple feedback on performance. It has been found that learner drivers may prefer to increase speed rather than to drive more accurately to comply with the norms (De Winter et al., 2006d). This distinction between speed and accuracy seems to be similar to the distinction between violations and errors, defined here as intentional and unintentional deviations from normative performance respectively (see also Rothengatter, 1997). It has been suggested that errors could be counteracted by means of proper training (e.g., Reason et al., 1990), whereas violations could be better prevented by self-reflection and attitude enforcement (e.g., Hatakka et al., 2002). Because both types of behaviour require different modes of communication, a virtual instructor should be able to distinguish between intentional violations and unintentional errors. Therefore, objective methods are needed to make a distinction.

The violation-error distinction has been extensively studied by means of questionnaires. Reason et al. (1990) were the first to show, using the Driver Behaviour Questionnaire (DBQ), that driving behaviour on the roads can be divided into violation and error components. They suggested that violations and errors are caused by different psychological processes. Others have repeated and extended this work. For example, demographic differences and the relation of violations and errors to road accidents have been investigated (Åberg & Rimmö, 1998; Blockey & Hartley, 1995; Kontogiannis et al., 2002; Lajunen et al., 2004; Mesken et al., 2002; Özkan & Lajunen, 2005; Parker et al., 1998; Rimmö & Åberg, 1999; Xie & Parker, 2002).

Although violations and errors are probably governed by different psychological processes, this does not imply that both types of aberration are unrelated. There are numerous indications that drivers show adaptive behaviour. For example, it has been reported that skid control training indeed improves vehicle-handling skills but that it also leads to an increased number of speed violations (Katila et al., 1996), and Van Winsum and Godthelp (1996) showed that drivers with better steering skills adopt higher speeds in bends. Furthermore, speed control plays an important role in driving behaviour because speed influences task conditions and hence the susceptibility to violations or errors (Rothengatter, 1997; Fuller, 2005). Without doubt, violations, errors, and speed are intimately related and further investigation is therefore needed to support the development of virtual instructors.

Distinguishing between violations and errors without completing questionnaires is easier said than done. First, a computer cannot assess a person's intentions directly, although progress is made in correlating physiological measurements with driving performance (e.g., Lin et al., 2005). Second, there exists no acceptable definition of normative driving behaviour (Kappé & Van Emmerik, 2005; Pirenne et al., 2002; Rothengatter, 1997). Traffic law regulations do not suffice, because they implicitly assume intentional behaviour. Driving through a red light, for example, is always regarded as a violation, even if the driver erroneously failed to see the light. Third, during an action sequence, violations and errors can occur simultaneously (Reason et al., 1990) and therefore observed driving behaviour can be the result of both a violation and an error.

These difficulties also play a role in the development of driver support systems (Rothengatter, 1997). Rothengatter (1991) proposed to simplify the problem by attributing every deviation from normative performance to intentional violations. Another option is "to consider deviations from normative performance as intentional only when these 'normally' occur with specific drivers" (Rothengatter, 1997). In other words, given the fact that a group of violation-prone drivers (for example, men) is observed to frequently deviate from particular normative performance (for example, not keeping enough distance), this deviation can generally be considered a violation. Similarly, Reason et al. (1990) stated that: "Distinctions between errors and violations, on the one hand, and their respective contributions to road accidents, on the other, emerge more clearly from group analyses in which age and sex differences, rather than individuals, are the focus of study" (p. 1317). Gender differences have been identified in DBQ scores, with men reporting more violations than women and women reporting more errors than men (Åberg & Rimmö, 1998; Blockey & Hartley, 1995; Parker et al., 1995; Reason et al., 1990).

The aim of this study was to find out whether violations and errors could be identified in objective performance data. For this purpose, performance data of participants who completed a simulation-based driver training programme were analysed and the relationships between violations, errors, and speed were assessed.

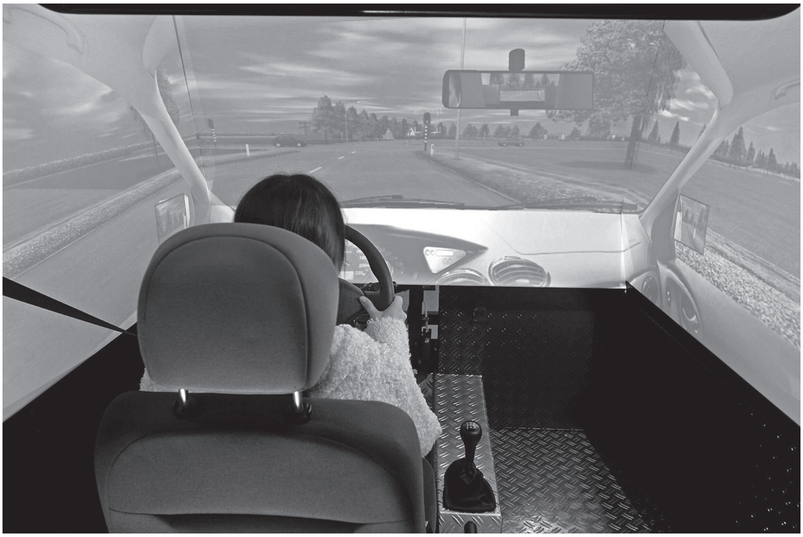


Figure 1. Dutch Driving Simulator (Green Dino, 2007).

2. Method

For a period of 8 months, performance data were recorded for participants involved in simulation-based driver training in simulators placed throughout the Netherlands.

2.1. Hardware

The driving simulators used in this study were the commercially available fixed-base Dutch Driving Simulators (DDS; Green Dino, 2007) (Figure 1). The vehicle controls of the DDS were similar to an actual car with a manual transmission. Force feedback was provided on the steering wheel according to the self-aligning torque of the front wheels and acceleration cues were provided by means of vibrations on the steering wheel column and vibrations in the back of the seat. Three projectors provided a geometrical 180° field of view. The resolutions were 1024 x 768 pixels for the front view projection and 800 x 600 pixels for the side view projections. The dashboard, the vehicle interior, and mirrors were integrated in the projected image.

2.2. Training software

The curriculum was based on Dutch driver training and consisted of 15 lessons, 27 min each. Lessons 1–5 were dedicated to vehicle control, lessons 6–10 to driving in urban areas with intersections and roundabouts, and lessons 11–15 to motorway driving. A human supervisor was able to manually alter the default order of the lessons. Each lesson consisted of two or three sessions with the duration ranging from 4 min 30 s to 16 min 4 s. Sessions were preceded by instructive text, voice, and movies. In total, the training programme consisted of 1 h of such multimedia instructions and 5 h 45 min of simulation-based driving.

A virtual instructor provided instructions and feedback on performance. After the task was successfully executed a successive number of times, instructions and feedback were reduced. Tasks were automatically activated and deactivated according to predefined conditions. For example, the task of taking bends was activated when entering the bend and deactivated when leaving the bend. Tasks could also end as a result of a 30-s time-out. Different tasks could be activated within a session, but only one task could be activated at a time. Tasks were automatically assessed by a virtual instructor that monitored whether the participant exceeded predefined (normative) thresholds. Most tasks could be failed for different reasons. More information about the virtual instructor can be found in Weevers et al. (2003a, 2003b).

2.3. Participants

A total of 42 different driving schools (totalling 46 DDSs) across the Netherlands took part in the measurements. Performance data of 2,530 participants were recorded, 520 of whom had completed all lessons in default order (263 men, 257 women, mean age based on available records =19.7 years). As these participants had completed the same training programme, their data were used for the analysis that is presented in this chapter. The median number of days needed to finish the default training programme was 17.5. After having completed the simulation-based training programme, participants continued training in a real vehicle.

2.4. Measures

Driving performance on 20 tasks (see Table 1) was automatically recorded. All data were brought together after completion of the 8-month measurement. The number of successes, the number of failures per failure reason, and the mean task completion time (MTCT) were calculated for every participant for every task. If a participant had failed a task for more than one reason, then only the failure reason listed first in Table 1 was recorded. All time-outs were excluded from the analysis. Only task successes were used to calculate MTCT; the MTCT was not calculated when the participant was never successful on a task. The number of task activations for a participant-task combination was defined as the number of successes plus the number of failures.

Table 1. Means and standard deviations of the number of successes for each task (T1 to T20), the mean task completion time (MTCT) for each task, and the number of failures for each failure reason (F1 to F52) (Note that failure reasons belong to task)

		Successes		MTCT				Failures	
Task		Mean	SD	Mean	SD		Failure reason	Mean	SD
T1	Driving away during start/stop exercise [a]	12.79	4.64	15.32	4.95	F1	Stalling the engine	0.06	0.34
						F2	Throttle > 75% when starting engine	0.09	0.38
						F3	Wrong pedal or wrong gear	0.17	0.47
						F4	Clutch pedal speed > 100% per s	2.20	2.26
						F5	Throttle > 75% when in gear	0.12	0.48
T2	Driving away from parking lane	1.42	1.36	31.01	11.76	F6	Stalling the engine	1.54	2.03
						F7	Wrong pedal or wrong gear	0.17	0.46
						F8	No indicator during lane change	0.77	1.07
						F9	Lateral speed > 0.5 lane per s	0.09	0.35
T3	Driving on and near intersections	82.35	11.57	38.63	3.66	F10	Collision with other car, pedestrian, or bicyclist	0.98	1.13
						F11	Driving too fast with respect to others having right of way	16.88	3.98
						F12	Driving too fast (no need to stop)	12.62	7.45
						F13	Approaching red traffic light too fast	1.50	1.54
						F14	Pressing clutch when decelerating and having to turn	6.22	4.81
						F15	Not in 2nd gear when having to turn or near other car(s)	2.30	3.57
						F16	Approaching amber traffic light too fast	1.95	1.37

(*table continues*)

Table 1. (continued)

Task		Successes		MTCT				Failures	
		Mean	SD	Mean	SD		Failure reason	Mean	SD
T4	Keeping maximum speed in bends	17.06	4.33	17.31	1.98	F17	Driving more than 5 km/h too fast	3.46	3.18
T5	Keeping maximum speed on motorway segments	52.42	10.55	20.13	6.86	F18	Driving more than 5 km/h too fast	16.56	7.98
T6	Keeping safe distance in intersections world	0.59	1.04	10.36	3.01	F19	Time headway < 1 s	0.22	0.60
T7	Keeping safe distance on winding road	1.69	2.73	11.13	6.39	F20	Time headway < 1 s	0.90	2.07
T8	Keeping safe distance on motorways	10.85	6.49	14.00	3.21	F21	Time headway < 1 s	8.34	6.78
T9	Lateral motorway manoeuvres	171.47	31.53	15.84	2.23	F22	No indicator at exit lane or during lane change	27.54	19.74
						F23	Wrong pedal or wrong gear	0.14	0.43
						F24	Lateral speed greater than half a lane per s	4.72	6.15
T10	Lane tracking[b]	4.10	2.08	29.04	1.96	F25	Lane centre error > 1.40 m	1.49	1.43
						F26	Lane centre error more than 2 s between 0.85 m and 1.40 m	0.62	0.92
						F27	Pressing clutch when decelerating in turn	0.55	0.88
						F28	Lane centre error less than 2 s between 0.85 m and 1.40 m	2.09	1.47

(*table continues*)

Table 1. (continued)

Task		Successes		MTCT				Failures	
		Mean	SD	Mean	SD		Failure reason	Mean	SD
T11	Shifting gears	53.33	13.20	9.81	4.26	F29	Pressing wrong pedal/shifting to wrong gear	0.06	0.29
						F30	Clutch pedal speed > 200% per s	13.94	9.83
						F31	Throttle > 75% when in gear	0.83	1.60
T12	Shifting gears during shifting gears exercise [a,b]	13.11	4.28	9.70	3.11	F32	Wrong pedal or wrong gear	0.59	0.82
						F33	Clutch pedal speed > 200% per s	0.65	1.17
						F34	Throttle position > 75% when in gear	0.16	0.50
T13	Speed tracking [a,c,d]	10.61	2.50	17.73	0.32	F35	Deviating more than 5 km/h from instructed speed	1.60	1.75
T14	Spotting objects (spatial-perceptual task) [a,b,d]	18.14	1.57	18.83	0.88	F36	Not pressing horn before passing the object (too late)	0.90	1.12
						F37	Pressing horn while object is not yet visible (too early)	0.40	0.77
T15	Stopping during start/stop exercise [a]	16.67	4.30	15.84	6.23	F38	Stalling the engine	0.47	2.10
T16	Stopping in front of stop sign	2.49	1.69	39.40	11.44	F39	Stalling the engine	0.16	0.41
						F40	Brake pedal speed > 1500% per s	0.11	0.34
						F41	Pressing clutch when engine speed > 1200 rpm	1.45	1.30
						F42	Not pressing clutch so that engine almost stalls	1.62	1.58

(*table continues*)

Table 1. (continued)

Task		Successes		MTCT				Failures	
		Mean	SD	Mean	SD		Failure reason	Mean	SD
T17	Stopping the car in parking lane	3.45	1.51	40.21	10.38	F43	Driving on soft shoulder	0.01	0.11
						F44	No indicator during lane change	0.76	0.98
						F45	Lateral speed > 0.5 lane per s	0.15	0.52
T18	Taking bends during exercise	5.64	2.57	10.39	3.83	F46	Lane centre error > 1.0 m and increasing more than 0.3 s	3.07	2.21
						F47	Driving faster than 30/60 km/h in sharp/normal turns	2.08	1.58
						F48	Pressing clutch when decelerating in turn	1.18	1.48
T19	Taking bends during exercise [b]	5.94	2.81	7.02	1.91	F49	Lane centre error > 1.0 m and increasing more than 0.3 s	5.68	2.79
T20	Taking exit and entry ramps	21.73	11.23	17.71	5.30	F50	Lane centre error > 1.0 m and increasing more than 0.3 s	19.05	8.56
						F51	Driving faster than 30/60 km/h in sharp/normal turns	5.37	3.76
						F52	Pressing clutch when decelerating in turn	0.73	1.11

[a] Automated steering
[b] Automated speed of the car
[c] Reactivated at fixed times
[d] Automated gear changing

3. Results

3.1. Frequencies

Table 1 shows the means and standard deviations of the number of failures per failure reason, the number of successes per task, and the MTCT per task. It can be seen that the mean number of successes and mean number of failure reasons differed highly between tasks, which can be explained by the curriculum design and the strictness of the normative performance. For example, lateral highway manoeuvres (T9) was activated many times, which can be explained by the fact that multiple sessions were devoted to lane changes and overtaking. Stopping in front of a stop sign (T16) is an example of a task that was activated less frequently, because it was given in one short-lasting session only and the mean MTCT was relatively high. Because the assessment criteria of this task were rather strict, on average, more failures (3.34 which is the sum of failure reasons F39, F40, F41, and F42) than successes (2.49) were counted.

An interesting finding was that the average man had more failures than the average woman (184 vs. 167, Cohen's *d* effect size (*d*) = 0.37, $p < 0.001$ using a *t* test) and more successes as well (521 vs. 491, $d = 0.56$, $p < 0.001$). In other words, men were involved in more task activations, even though every participant received equal training time.

3.2. Factor analysis of failure reasons

3.2.1. Factor analysis

To investigate whether a distinction can be made between violations and errors, the Pearson correlation matrix among the failure reasons was constructed and factors were extracted using the maximum likelihood method. Only two factors were extracted, because the violation-error distinction was of interest. The scree plot clearly supported this two-factor solution. Parallel analysis (O'Connor, 2000) indicated that an eight-factor solution would be most appropriate. However, the eight-factor solution was not readily interpretable and was clearly an overextraction resulting from the fact that parallel analysis is sensitive to sample size thereby recommending an inappropriately high number of factors. Oblique rotation (oblimin) was tried and it was found that only a small negative correlation (-0.10) existed between the factors, so it was decided to simplify the solution by orthogonal Varimax rotation. The result of the factor analysis is shown in Table 2. The two orthogonal factors accounted for 11.9% of the total variance: 7.35% for factor 1 and 4.52% for factor 2. The small variance can be explained by several failure reasons that hardly contributed to the overall solution. After removing 34 failure reasons having a correlation with all other failure reasons that was smaller than 0.30 (Tabachnick & Fidell, 2001), the two-factor solution would explain 25.3% of the variance, which is comparable to the DBQ-

Table 2. Factor loadings and communalities (h^2) of failure reasons in the violation-error solution

Failure reason	Violations	Errors	h^2
F1 Stalling the engine	0.06	0.09	0.01
F2 Throttle > 75% when starting engine	0.00	0.03	0.00
F3 Wrong pedal or wrong gear	0.02	0.23	0.05
F4 Clutch pedal speed > 100% per s	0.28	0.04	0.08
F5 Throttle > 75% when in gear	0.00	0.08	0.01
F6 Stalling the engine	0.01	0.20	0.04
F7 Wrong pedal or wrong gear	-0.06	0.22	0.05
F8 No indicator during lane change	0.11	0.02	0.01
F9 Lateral speed > 0.5 lane per s	-0.01	0.14	0.02
F10 Collision with other car, pedestrian, or bicyclist	0.13	0.04	0.02
F11 Driving too fast with respect to others having right of way	0.29	-0.08	0.09
F12 Driving too fast (no need to stop)	0.43	0.28	0.26
F13 Approaching red traffic light too fast	0.09	0.24	0.07
F14 Pressing clutch when decelerating and having to turn	0.39	-0.25	0.21
F15 Not in second gear when having to turn or near other car(s)	0.07	0.16	0.03
F16 Approaching red traffic light too fast	-0.14	0.17	0.05
F17 Driving more than 5 km/h too fast	0.57	-0.17	0.35
F18 Driving more than 5 km/h too fast	0.72	0.06	0.53
F19 Time headway < 1 s	0.46	0.18	0.25
F20 Time headway < 1 s	0.44	-0.03	0.19
F21 Time headway < 1 s	0.62	-0.01	0.39
F22 No indicator at exit lane or during lane change	0.27	0.07	0.08
F23 Wrong pedal or wrong gear	-0.01	0.11	0.01
F24 Lateral speed greater than half a lane per s	0.19	0.29	0.12
F25 Lane centre error > 1.40 m	0.01	0.45	0.20
F26 Lane centre error more than 2 s between 0.85 m and 1.40 m	-0.10	0.32	0.11
F27 Pressing clutch when decelerating in turn	0.29	-0.46	0.30
F28 Lane centre error less than 2 s between 0.85 m and 1.40 m	0.06	0.21	0.05
F29 Pressing wrong pedal/shifting to wrong gear	0.00	0.06	0.00
F30 Clutch pedal speed > 200% per s	0.68	-0.09	0.47
F31 Throttle > 75% when in gear	0.25	0.09	0.07
F32 Wrong pedal or wrong gear	-0.12	0.23	0.07
F33 Clutch pedal speed > 200% per s	0.37	-0.09	0.15
F34 Throttle position > 75% when in gear	0.21	-0.05	0.05

(*table continues*)

Table 2. (continued)

Failure reason		Violations	Errors	h^2
F35	Deviating more than 5 km/h from instructed speed	-0.05	0.32	0.10
F36	Not pressing horn before passing the object (too late)	-0.04	0.07	0.01
F37	Pressing horn while object is not yet visible (too early)	0.04	-0.01	0.00
F38	Stalling the engine	0.06	0.22	0.05
F39	Stalling the engine	-0.04	0.12	0.02
F40	Brake pedal speed > 1500% per s	0.13	-0.05	0.02
F41	Pressing clutch when engine speed > 1200 rpm	0.39	-0.17	0.18
F42	Not pressing clutch so that engine almost stalls	-0.29	0.34	0.20
F43	Driving on soft shoulder	-0.05	0.09	0.01
F44	No indicator during lane change	0.04	-0.01	0.00
F45	Lateral speed > 0.5 lane per s	0.27	-0.05	0.07
F46	Lane centre error > 1.0 m and increasing more than 0.3 s	-0.09	0.52	0.28
F47	Driving faster than 30/60 km/h in sharp/normal turns	0.18	0.05	0.04
F48	Pressing clutch when decelerating in turn	0.38	-0.31	0.24
F49	Lane centre error > 1.0 m and increasing more than 0.3 s	-0.13	0.59	0.37
F50	Lane centre error > 1.0 m and increasing more than 0.3 s	0.05	0.24	0.06
F51	Driving faster than 30/60 km/h in sharp/normal turns	0.36	-0.01	0.13
F52	Pressing clutch when decelerating in turn	0.05	-0.01	0.00
	Eigenvalues and sum of squares	4.76	2.97	6.17
	Percentage explained	7.35	4.52	11.87

study of Blockey and Hartley (1995). Failure reasons loading above 0.20 were considered high enough (salient) to assume that a relationship existed between the failure reason and the factor (Gorsuch, 1974, taking into account the sample size of 520).

3.2.2. Reliability analysis

The internal consistencies (Cronbach's α) of salient variables were 0.79 for factor 1 and 0.65 for factor 2, which seems to be fairly low but around the same level as those found after factor analysing DBQ-responses (Lajunen et al., 2004).[1] Additionally, the reliability of the factor loadings was assessed by randomly dividing the 520

[1]) Note that a high Cronbach's α is not a necessary condition for a good factor analysis (Boyle, 1991; appendix B). Instead, maximizing Cronbach's α can result in overly specific factors with low validity.

participants into two groups of 260 participants and reperforming the factor analysis for these two subsamples. Factorial agreement was then assessed by the congruence coefficient (Levine, 1977). The factor loading congruence coefficients between the subsamples and the total sample were 0.94 and 0.98 and the congruence coefficient between both subsamples was 0.87, indicating that reasonably equal results could be produced by two independent (and smaller) groups of participants.

3.2.3. Factor 1 description

Table 2 shows that failure reasons loading (> 0.20) on factor 1 were almost exclusively excess of speed and the breaching of safety-related thresholds. The failure reasons loading most highly (> 0.40) on this factor were: driving too fast (F12, F17, F18); moving the clutch too quickly (F30); and following too closely (F19, F20, F21). Not using the traffic indicator (F22) and pressing the clutch when it was forbidden by the rules (F14, F27, F41, F48) also loaded on the first factor. The latter procedure followed Dutch training regulations, which state that drivers should be in the desired gear before executing a manoeuvre to prevent excessive increase of task difficulty (Vissers, 2001).

3.2.4. Factor 2 description

Failure reasons that loaded (> 0.20) on factor 2 were related to poor vehicle control. The highest factor 2 loadings (> 0.30) represented lane-tracking or speed-tracking errors (F25, F26, F35, F46, F49). Lower loadings generally related to procedural failures such as stalling the engine and shifting to the wrong gear or applying the wrong pedal. Interestingly, approaching a red or amber traffic light too fast (F13, F16) and driving too fast on and near intersections (F12) loaded moderately on factor 2 as well, although the characteristics of these failure reasons resemble the characteristics of factor 1. Personal observations can explain the loadings on factor 2. Traffic lights and speed signs were relatively difficult to perceive (see also Figure 1) and often disappeared behind the projected rear view mirror, so these aberrations can be explained as the result of a perceptual deficiency in the driving simulator.

3.2.5. Negative factor loadings

Negative factor loadings have a distinct meaning as well. While positive factor 1 loadings were found for speeding and breaching safety-related thresholds, negative factor 1 loadings may represent a reluctance to act in time. Not pressing the clutch so that the engine almost stalls (F42) and approaching an amber traffic light too fast (F16) both belong to this category. While positive factor 2 loadings were related to poor vehicle control, negative loadings appear to be a sign of good vehicle control and especially occurred for applying the clutch when against the rules

(F14, F27, F41, F48). It can be reasoned that participants with better vehicle-handling abilities were involved in these failures because they were less obstructed by increased task difficulty. For example, the Pearson correlation between F27 and F49 is -0.29 ($p < 0.001$) indicating that participants who had fewer lane-tracking failures more often applied the clutch in a bend.

3.2.6. Comparison to DBQ and gender differences

When comparing the factor distinction to the violation-error distinction from factor analyses of DBQ responses, it appeared that both categories are quite similar. Behaviours such as driving too close and speeding were clustered into the violation category, whereas perceptual, judgemental, and vehicle control-related deficiencies were clustered into the error categories (e.g., Reason et al., 1990).

This study supports the finding that the two factors represent violations and errors. Studies using a DBQ and a factor analysis have shown that men are generally more involved in violations and that women are more involved in errors (e.g., Reason et al., 1990). Figure 2 shows gender differences in terms of effect size (Cohen's *d*) versus factor loading differences (factor 1 loading minus factor 2 loading) for each of the 52 failure reasons. It can be seen that a clear relationship exists. When

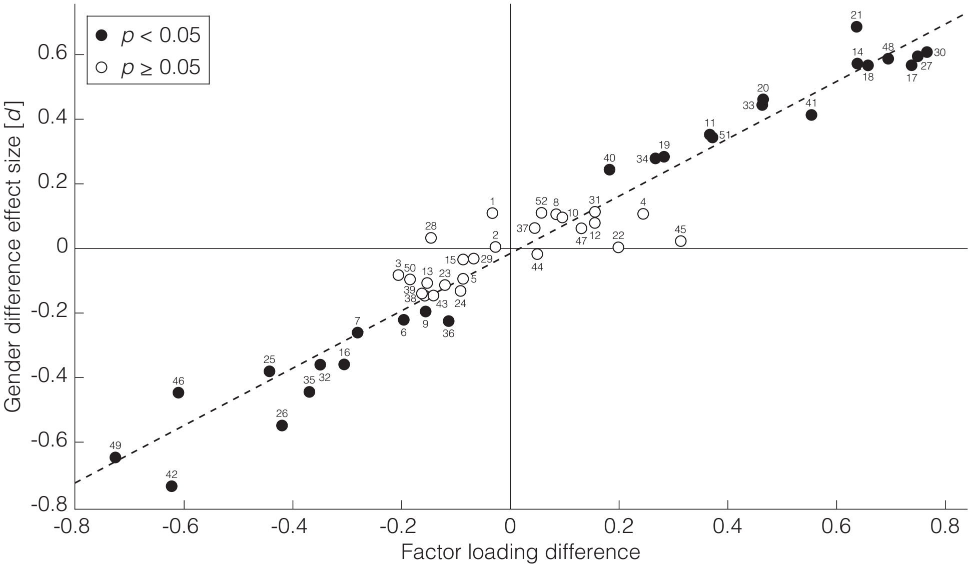


Figure 2. Magnitude of gender difference (Cohen's *d* effect size) of number of failures versus factor loading difference (violation factor loading minus error factor loading) for each of the 52 failure reasons.

factor loading difference is high (failure reason is more a violation than an error), men had more failures; when the factor loading difference is low (failure reason is more an error than a violation), women had more failures. A Pearson correlation between gender differences and factor loading differences supported this observation ($r = 0.97$, $p < 0.001$). Pressing the horn too late on the spatial-perceptual task (F36) occurred significantly more amongst women, which is in accordance with literature about gender differences in spatial abilities (e.g., Coluccia & Louse, 2004). The correspondences between factor loadings and gender differences supports that factor 1 represents violations and factor 2 represents errors.

3.3. Factor analysis of mean task completion times

3.3.1. Factor analysis

To investigate the speed choice of participants, a Pearson correlation matrix of the MTCTs of the 20 tasks was calculated; missing items were excluded pairwise. Parallel analysis indicated that two factors should be extracted; however, the two-factor solution was not regarded as interpretive. The scree plot clearly indicated that a one-factor solution was most appropriate, so it was decided to extract only one factor using the maximum likelihood method. The factor accounted for 15.6% of the total variance. The factor loadings are shown in Table 3.

3.3.2. Reliability analysis

The internal consistency (Cronbach's α) of the salient variables was 0.78. The factor loading congruence between two random subsamples of 260 participants was 0.96 and the factor congruences between each subsample and the total sample were 0.99 each, indicating that the factor loadings can be considered reliable.

3.3.3. Factor description

The extracted factor can best be described as the extent to which motivation for speed was expressed in the MTCT and is referred to as the paceability [2] of the driving task. The highest factor loadings (> 0.50) were found for tasks for which the participant was relatively unconstrained concerning the task completion time. For example, it can be imagined that the time taken to complete straight road segments (T5) or to complete lateral highway manoeuvres (T9) depended largely on the participant's preferred speed. Taking bends (T4, T18, T20), stopping the car (T15, T16, T17), and driving away from a parking lane (T2) had lower factor loadings, because the MTCT was more ambiguously related to motivation for speed. External factors, such as the task difficulty, relation to initial speed or the predefined criteria that

[2]) In other chapters, the factor extracted from the task completion times is referred to as the "speed factor"

Table 3. Factor loadings and communalities (h^2) of tasks in the paceability solution

Task		Paceability	h^2
T1	Driving away during start/stop exercise [a]	0.41	0.17
T2	Driving away from parking lane	0.09	0.01
T3	Driving on and near intersections	0.65	0.43
T4	Keeping maximum speed in bends	0.41	0.17
T5	Keeping maximum speed on motorway segments	0.77	0.60
T6	Keeping safe distance in intersections world	-0.16	0.02
T7	Keeping safe distance on winding road	-0.27	0.07
T8	Keeping safe distance on motorways	-0.08	0.01
T9	Lateral highway manoeuvres	0.68	0.46
T10	Lane tracking [b]	-0.13	0.02
T11	Shifting gears	0.56	0.31
T12	Shifting gears during shifting gears exercise [a,b]	0.56	0.31
T13	Speed tracking [a,c,d]	-0.08	0.01
T14	Spotting objects (spatial-perceptual task) [a,b,d]	0.05	0.00
T15	Stopping during start/stop exercise [a]	0.33	0.11
T16	Stopping in front of stop sign	0.24	0.06
T17	Stopping the car in parking lane	0.42	0.18
T18	Taking bends during exercise	0.18	0.03
T19	Taking bends during exercise [b]	0.02	0.00
T20	Taking exit and entry ramps	0.39	0.15
	Eigenvalue and sum of squares	3.82	3.11
	Percentage explained		15.57

[a] Automated steering

[b] Automated speed of the car

[c] Reactivated at fixed times

[d] Automated gear changing

started and ended the tasks, played a more important role here, so that the MTCT was less informative about the participant's preferred speed. Tasks for which the MTCT or the speed were computer-controlled (forced-paced tasks T10, T13, T19) had near-zero factor loadings. The distance-keeping tasks (T6, T7, T8) showed negative factor loadings, because these tasks were active when driving within 2 s distance of another car. Generally, drivers with a tendency to be quick followed the car in front more closely so that the task remained active for a longer time, so, paradoxically, participants who wanted to be quick took longer to complete distance-keeping tasks.

3.3.4. Comparison to gender differences

Again, the extracted factor was compared with gender differences. Figure 3 shows that a clear relationship exists between the paceability factor and the effect size of the MTCT of men and women. The larger the factor loading, the lower the effect size (r = -0.96, p < 0.001), indicating that men have a higher inclination for speed than women.

3.4. Factor score correlations

Violation-scores, error-scores, and speed-scores were calculated for each participant according to the Bartlett weighted least squares procedure. The sign of the speed-score was reversed so that increased motivation for speed corresponded to an increase in speed-score. Additionally, activations-scores were calculated from the mean z-transformed number of task activations; the activation-scores are an indicator for the number of tasks completed by the participant.

Figure 4 shows the activations-score versus the speed-score for each participant. It can be seen that participants having a high motivation for speed were in-

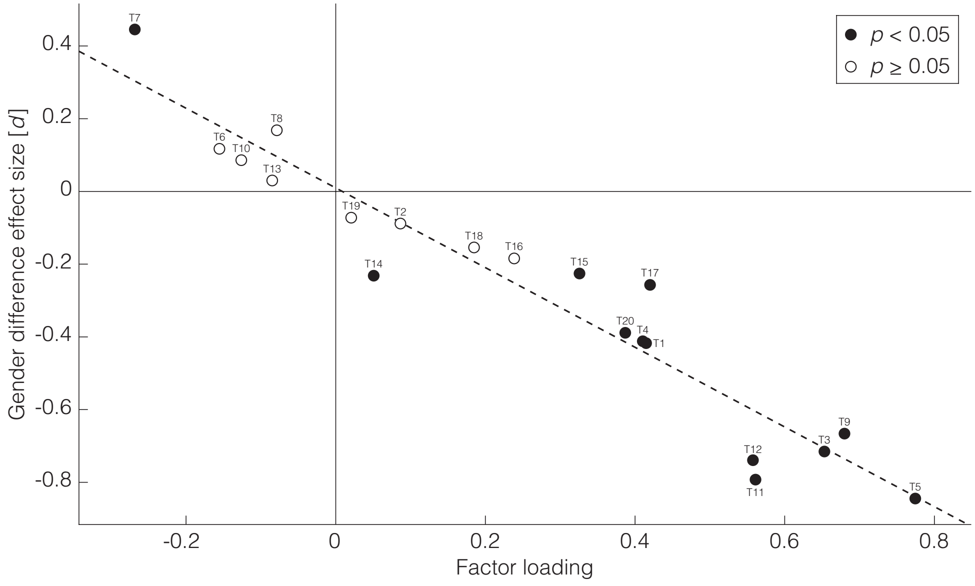


Figure 3. Magnitude of gender difference (Cohen's *d* effect size) of mean task completion time versus paceability factor-loading for each of the 20 tasks.

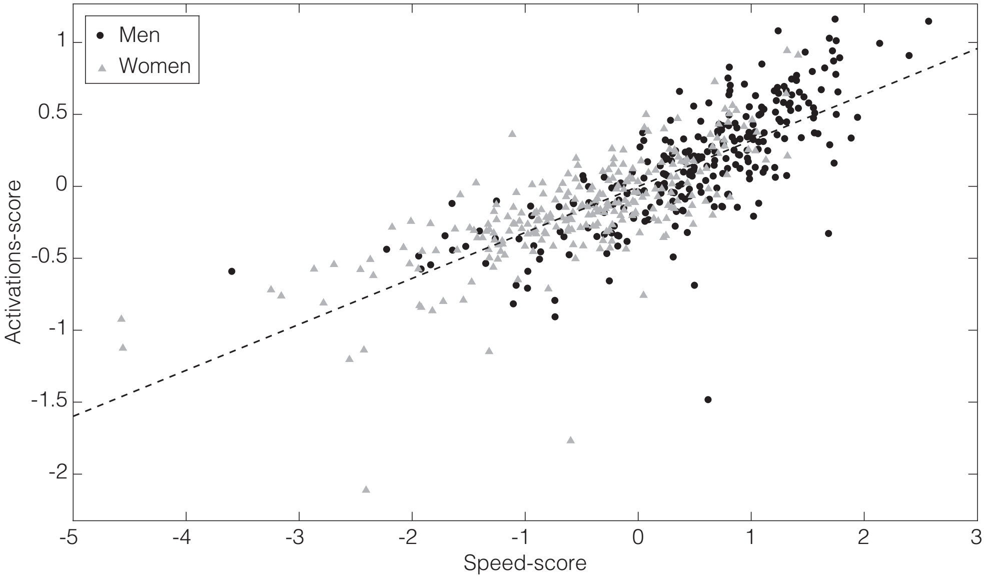


Figure 4. Activations-score versus speed-score for each of the 520 participants. The dashed line is a linear fit.

volved in more tasks (*r* = 0.78, *p* < 0.001). On average, men had a higher speed-score than women (0.46 vs. -0.48, *d* = 1.06, *p* < 0.001) and a higher activations-score as well (0.36 vs. -0.37, *d* = 0.78, *p* < 0.001). The latter result corresponds to the aforementioned result that men had both more successes and more failures than women (section 3.1).

Figure 5 shows the violation-score versus the speed-score. It can be seen that speed and violations bear a strong relationship (*r* = 0.71, *p* < 0.001) insofar as quicker participants committed more violations. This makes sense when considering that participants with a higher speed-score were involved in more tasks (see Figure 4) and that many violations were found to be speed related. It can also be seen that participants with a very high speed-score (> 1.50) were involved in an increased number of violations. On average, men had a higher violation-score than women (0.38 vs. -0.39, *d* = 0.83, *p* < 0.001).

Figure 6 shows that motivation for speed had a small inverse correlation with errors (*r* = -0.24, *p* < 0.001). Men had a lower mean error-score than women (-0.37 vs. 0.38, *d* = -0.81, *p* < 0.001).

3.5. Change of behaviour with time

To further investigate the relationships between errors, violations and speed, the participant's temporal behaviour within sessions was addressed. A distinction was

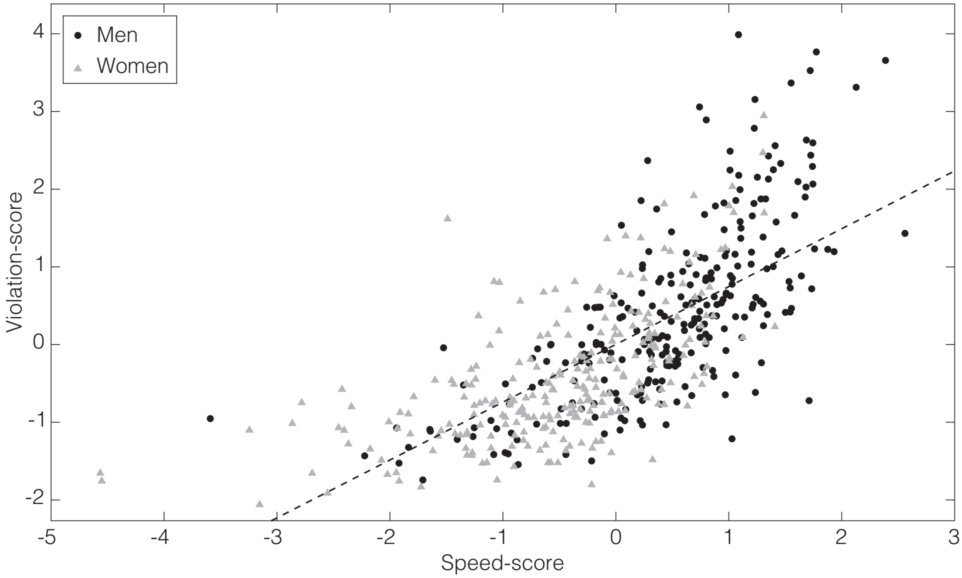


Figure 5. Violation-score versus speed-score for each of the 520 participants. The dashed line is a linear fit.

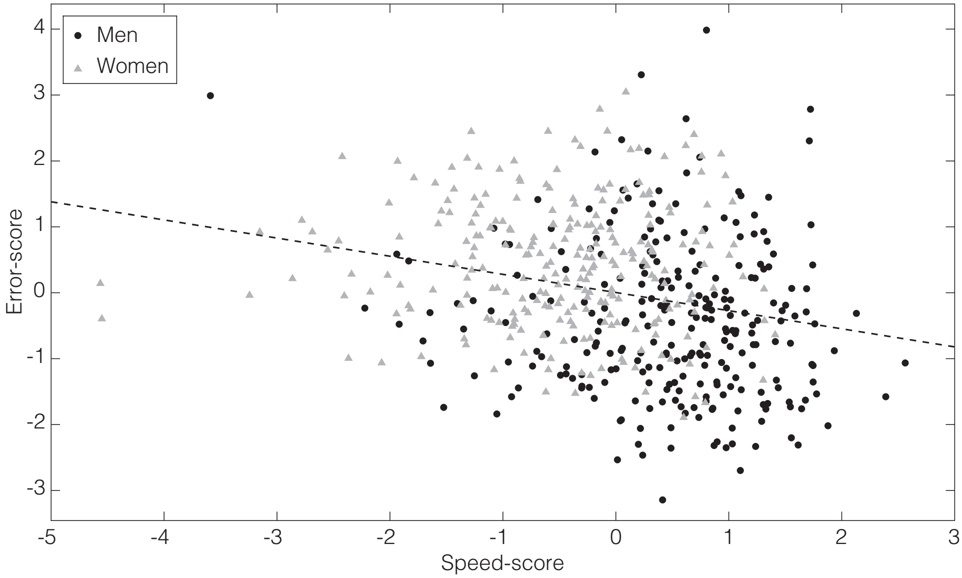


Figure 6. Error-score versus speed-score for each of the 520 participants. The dashed line is a linear fit.

made between forced-paced (paceability near zero) and self-paced tasks (paceability considerably greater than zero).

3.5.1. Forced-paced

For forced-paced tasks, plotting the mean cumulative number of failures and task activations versus time typically yields a graph as shown in Figure 7. This figure presents the task lateral position tracking (T10), during which participants had to drive accurately in the middle of the right lane of a winding road. The waving pattern occurred because the task was reactivated every 30 s. The solid lines represent the mean cumulative number of two failure reasons (F25 + F26). Only these two failure reasons were shown because they could be characterized as errors (see Table 2). It can be seen that the slope of these curves decreases with time, which can be explained by performance improvement. The dotted lines show that the mean number of errors can well be described by means of a simple experience curve as defined in equations (1) to (3).

$$f(t) = F0 \cdot t^b \quad (1)$$

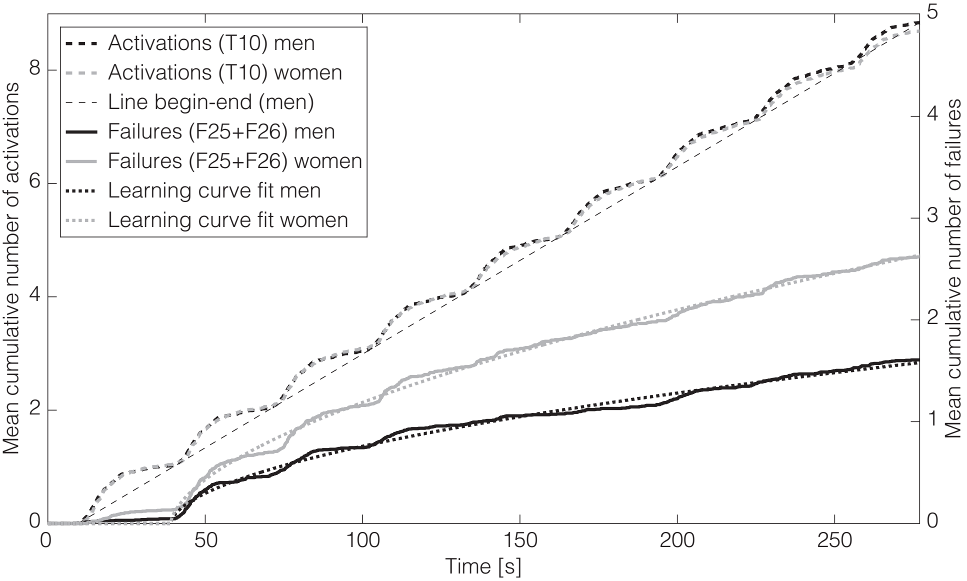


Figure 7. Mean cumulative number of task activations (task T10) and mean cumulative number of failures (failure reasons F25 and F26) versus time in the session. The thin black line through begin and endpoint was included to illustrate that the slopes of the curves remained constant with time.

Integration to obtain cumulative values yields:

$$F(t) = \frac{F0 \cdot t^{b+1}}{b+1} \tag{2}$$

In which t is the time since a starting moment, $f(t)$ is the failures per s at time t, $F0$ is the failures per s at $t = 0$, $F(t)$ is the cumulative number of failures at time t, and b is the learning constant.

The learning percentage is defined as:

$$\theta = 2^{b} \tag{3}$$

Best fits for men and women were obtained with reasonably similar learning percentages (θ) of 72.2 and 74.6 and with different begin values ($F0$) of 0.048 and 0.065 respectively. This indicates that men and women did not markedly differ in their learning abilities; however, they started at different initial performance levels. These similarities and differences respectively were also observed in other forced-paced tasks.

3.5.2. Self-paced

Figure 8 shows a similar graph but for the self-paced task driving away during start/stop exercise (T1). During the session, participants had to drive away repeatedly and bring the car to a stop for 10 min on an endless straight road. No other vehicles were present and participants did not have to steer. It can be seen that the slopes of the dashed grey and black curves increase, which can be explained by the fact that participants speeded up, hence increasing the task activation rate. Looking at the mean cumulative number of violations (F4, moving the clutch too quickly), it can be seen that the slopes of these lines also increase, indicating that (simulation-based) training actually led to an increased number of violations per time unit. Similar patterns were observed in other self-paced tasks.

4. Discussion

4.1. Distinction between violations and errors

Two factors were identified in the automatic recordings of failures in participants who completed a simulation-based driver training programme. This two-factor solution appeared to be similar to the violation-error distinction obtained from questionnaires, such as the DBQ, that asked respondents about intentional and unintentional driving behaviours. Although the main criterion of distinguishing between vio-

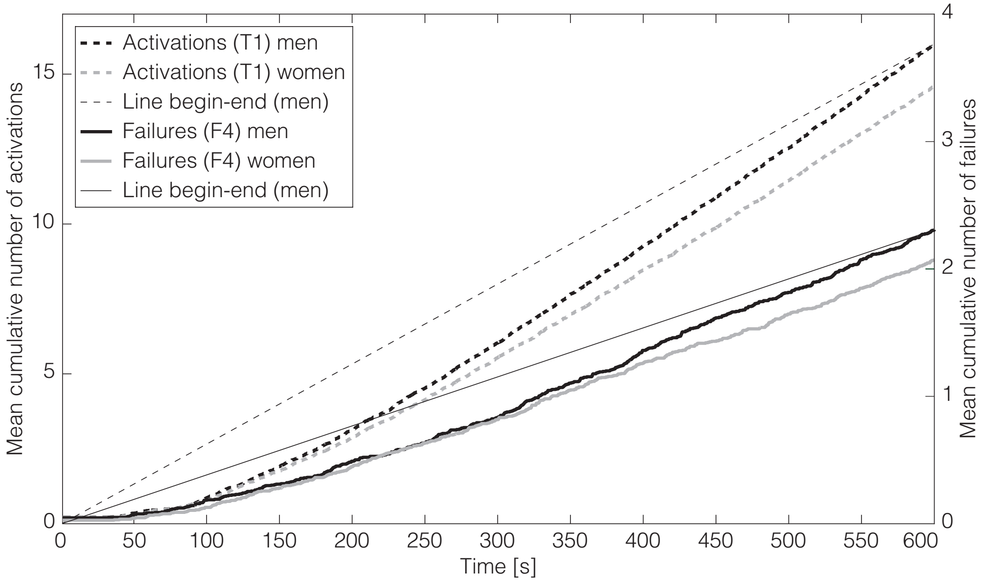


Figure 8. Mean cumulative number of task activations (task T1) and mean cumulative number of failures (failure reason F4) versus time in the session. The thin black lines through begin and endpoints were included to illustrate that the slopes of the curves increased with time.

lations and errors is intention, which was not measured in this study, the results support the finding that the extracted factors represented violations and errors. Factor loadings indicated that differences exist in the extent to which a failure can be regarded violation and error, and gender differences corresponded with these differences. Men committed more violations and women made more errors; the differences being the largest for the failure reasons that loaded highest on these scales. This is in accordance with DBQ-studies reporting that men are involved in more violations and fewer errors than women (e.g., Reason et al., 1990). Put differently, this study showed that it is possible to make a distinction between violations and errors without letting participants complete questionnaires; however, previous factor analysis results of questionnaire responses were needed as a basis.

4.2. Speed in relation to violations and errors

Factor analysis of the MTCTs revealed a factor called paceability, which can be described as to what extent the participant's inclination for speed translates into lower MTCTs. The paceability factor distinguishes between forced-paced tasks and self-paced tasks. During forced-paced tasks, participants could not (or hardly) influence task completion times because these were computer-controlled. Participants reduced errors during these tasks, and it was found that a power curve was useful to describe the number of errors over time. During self-paced tasks, participants generally increase their speed as well as the number of violations per time unit.

Speed, errors, violations, and the number of task activations (failures + successes) were closely related. Generally, quicker participants committed more violations, made fewer errors, and were involved in more tasks per time unit. It is remarkable that quicker participants made fewer errors, considering that quicker participants performed more tasks than slower participants and that quicker driving induces higher task demands (e.g., Fuller, 2005). Being quick can therefore be seen as a sign of having good vehicle-handling abilities. Gender differences in speed-scores and error-scores were rather large when considering that gender differences in mental abilities and choice reaction time are only small (Colom et al., 2000; Lorenz & Manzey, 2001). It is likely that the average male participant had more previous practice, for example, with controlling cars in computer games or otherwise.

4.3. Results in a broader perspective

By definition, a simulator does not provide a perfect representation of real driving. Safety has a different meaning and the consequence of committing violations is different compared to reality. Moreover, this study focused on a particular subset of people, namely youngsters around 19 years old who had the intention of obtaining their driving licence. Nonetheless, some results might be relevant to real driving behaviour on the roads. The results can clarify why violations cannot easily be prevented by practising (e.g., Hatakka et al., 2002) and why driver training cannot be considered an effective countermeasure to (speed-related) crashes (e.g., Mayhew & Simpson, 2002).

In a broader perspective, the results seem to relate to Fitts' law of human movement (Fitts, 1954). Although developed for much simpler tasks than car driving, Fitts' law states that the time needed to complete a movement depends on the level of accuracy and the human information-processing capabilities, or put differently, humans can use their capabilities to increase speed at the cost of accuracy or vice versa. In another study, such a speed-accuracy trade-off was identified with participants who practised lane-tracking on a driving simulator (De Winter et al., 2006d). The present study addressed group averages rather than the speed-accuracy trade-offs of individuals. When assuming that training on the simulator increases information-processing capabilities regarding the driving task, Fitts' law then predicts that, on average, training results in fewer errors (Figure 7) and lower task completion times (Figure 8). In a way, committing a violation is paradoxical, because violations mostly relate to the exceeding of speed-related thresholds. When considering that being quick is a sign of being competent, it requires self-control not to exceed the artificial boundary.

4.4. Implications for the development of virtual instructors

The present results can help in defining whether a failure reason is generally the result of intentional or unintentional behaviour. This knowledge can help to improve the effectiveness of virtual instructors, as it allows for task-specific feedback aimed at preventing violations or errors. A suggestion for rectifying violations in the simulator involves creating hazard awareness and establishing knowledge of traffic rules, for example, by means of small multimedia "campaigns". Errors might be prevented by putting emphasis on repeating the task and giving direct feedback on performance. In view of the fact that being quick is a sign of having good vehicle-handling abilities, it might be an interesting idea to develop training software that adapts itself to quick students so that they are automatically confronted with a more difficult task.

The violation-error structure explained only 11.9% of the total variance. Communalities of several failure reasons were low, indicating that these failures had large task-specific (unique) or random variance, unrelated to the skill for preventing errors or the tendency to commit violations. From this viewpoint, it could be considered whether tasks such as spotting objects (T14) should be removed from the curriculum. The unique variance of this task is probably not of interest, because real car driving does not involve horn pressing when objects appear and the number of failures hardly related to performance on other tasks. Second, it is recommended to research whether the assessment criteria can be improved so that the number of failures tells more about the driver. For example, stalling the engine (F1, F6, F38, F39) turned out to be a relatively rare event with low communalities and low mutual correlations, suggesting that it is something that can accidentally happen to anyone once in a while. Other criteria might be better able to distinguish whether someone is good in driving away and stopping the car or not.

Finally, driver assessment involves more than analysing whether one follows the norms. Because (simulation-based) driving is partially a self-paced task, the assessment software should not solely concentrate on the state of the vehicle with regard to predefined thresholds, but also on the driver's pace and the number of tasks. Instead of reporting the number of failures of a participant, the number of successes should be reported too. For the same reason, it is recommended to train not only in short-lasting scenarios, which is common practice in many training simulators. Instead, training should take place in longer-lasting sessions so that students can choose speed according to their own discretion.

Appendix 3A

This appendix provides additional information regarding the training software described in section 2.2.

Typically, the complexity of the lessons in the training programme increased progressively. So, eventually students should be able to handle complex situations and complete them without making errors or committing violations. For example, at the beginning of the first intersection lesson (lesson 6), participants had to repeatedly drive straight, turn left, and turn right, at unsignalised intersections without other traffic. Later intersection lessons (lessons 7 through 10) featured more complex tasks, including different types of signalised and sign-controlled intersections, roundabouts, and autonomous traffic.

In many of the lessons, there were other cars (agents) driving around autonomously; pedestrians and cyclists were present but did not move. Critical conflicts could occur with the agents, but there were no preprogrammed or triggered critical conflict scenarios or hazard perception scenarios such as in the work of others (e.g., Allen et al., 2007a). Rather, the lessons were simulations of real-world lessons with a human instructor. In most lessons, the routes through the virtual environment were not fixed; for example, when the participant turned into a different direction than instructed by the virtual instructor, the lesson continued as normal without restarting.

The recorded tasks clearly do not represent all of the student's actions during the 15-lesson training programme; this would be unfeasible because of inherent limitations of data storage facilities (see chapters 7-10 for experiments where all data were recorded on a frequency of 50 Hz, but with a smaller number of participants and for a few tasks only). The data in this chapter represent basic performance records (success or failure and task completion time) of primary tasks that were active during the lessons, aggregated across similar lessons and/or similar types of tasks. Chapter 4 includes a factor analysis that is similar to the factor analysis in chapter 3, but in chapter 4, all recorded data were used for each primary task and each lesson block separately.

It should be noted that it is likely that the final results (i.e., the factor scores) do not depend intimately on the functional capabilities of the Green Dino Simulator, because they represent a weighted average of normalised performance records of many diverse tasks. Whereas the score on each item separately depends strongly on its definition and type of assessment, the essence of the factor scores is that they aim to be generic. It is expected that the speed-score, for example, is a personal characteristic, and that it would correlate substantially with a speed-score extracted from a different type of simulator.

More information about the simulator and the virtual instructor can be found in Fikkert et al. (2006), Green Dino (2007), Kappé and Van Emmerik (2005), Weevers et al. (2003a, 2003b), as well as in chapters 4 and 5.

CHAPTER 4

Relationships between driving simulator performance and driving test results

Abstract

Simulators are being used to an increasing extent for driver training, allowing for the possibility of collecting objective data on driver proficiency under standardized conditions. However, relatively little is known about how learner drivers' simulator measures relate to on-road driving. This study proposes a theoretical framework that quantifies driver proficiency in terms of speed of task execution, violations, and errors. This study investigated the relationships between these three measures of learner drivers' (*N* = 804) proficiency during initial simulation-based training and the result of the driving test on the road, occurring an average of 6 months later. A higher chance of passing the driving test the first time was associated with making fewer steering errors on the simulator and could be predicted in regression analysis with a correlation of 0.18. Additionally, in accordance with the theoretical framework, a shorter duration of on-road training corresponded with faster task execution, fewer violations, and fewer steering errors (predictive correlation 0.45). It is recommended that researchers conduct more large-scale studies into the reliability and validity of simulator measures and on-road driving tests.

De Winter, J.C.F., De Groot, S., Mulder, M., Wieringa, P.A., Dankelman, J., & Mulder, J.A. (in press-a). Relationships between driving simulator performance and driving test results. *Ergonomics*. (adapted with minor textual changes)

1. Introduction

Various studies have attempted to improve driver training to reduce the large number of crashes among young drivers. The European BASIC project recommended that diverse training methods (e.g., professional training, accompanied driving, phasing the training) need to be applied together; one measure alone is not effective enough (Hatakka et al., 2003). In another EU-funded project, TRAINER, research was conducted into the use of driving simulators as a means to reduce road crashes (Dols et al., 2001). In view of the importance of research in driver training, it is considered relevant to further investigate the value of driving simulators. More specifically, the present study attempts to gain insight into how learner drivers' proficiency during simulation-based training relates to their results on the official Dutch driving licence test.

1.1. Data quality in simulation-based driver training

Simulators are being used to an increasing extent to train learner drivers. Advantages of a simulator are that goal-oriented virtual worlds and scenarios can be presented, and that driving performance can be assessed objectively and accurately in a standardized fashion (De Winter et al., 2006b; Wassink et al., 2006). It is impossible to incorporate these aspects into driver training on the road. In order to train learner drivers to negotiate crossroads, for example, they first have to drive to a suitable location. Weather conditions and the positions of other vehicles depend on coincidental factors, and the quality of the assessments depends on the inevitably subjective judgements of the human instructor.

One potential disadvantage of driving simulators is that, by definition, they provide only a representation of reality, not reality itself. It is a technical and psychological challenge to match simulator driving to the real situation. However, driving simulators have proved to be excellent instruments in studies where relative comparisons are important. Driving simulators have been used for research into central nervous system disorders, visual impairment, age, gender, driving experience, sleep apnoea, alertness, alcohol dose levels, the use of mobile telephones, and so forth (e.g., Gawron & Ranney, 1988; Lew et al., 2005; Reed & Green, 1999). Blaauw (1982) found that performance on a driving simulator gave more accurate distinction between experienced and inexperienced drivers than an assessment on the road. Boydstun et al. (1980) found a strong correlation (0.88) between performance on a simulator and a driving test on the road in a group of healthy participants and persons with perceptual motor handicap. H.C. Lee (2003) found that performance on a low-cost simulator could explain more than two thirds of the variance in on-the-road assessments amongst older drivers. Lew et al. (2005) concluded that automatic assessment of simulator performance in patients with traumatic brain injury provided valid measures that could be more sensitive predictors of future driving performance than traditional tests on the road with a human examiner. Thus, simulators offer potential advantages for objective data collection and assessing individual differences.

As far as driver training is concerned, it is obviously essential that the skills learned on the simulator can be transferred to the road, and that the performance data of the training programme carry some predictive meaning for future driving behaviour. Recent scientific studies have made optimistic conclusions regarding the potential effects of computer-based driver training on road safety (Allen et al., 2007a; Fisher et al., 2006). Still, relatively little is known about the quantitative relationships between learner drivers' performance variables measured in a simulator and those measured on-road.

1.2. Theoretical framework for modelling the driver

A simulator can produce huge amounts of data regarding an individual's driving performance. A theoretical framework is mandatory to process these data in order to obtain useful indicators of driver proficiency. Car driving is a very complex task, and no driver model can capture all of its intricacies (Reason et al., 1990). Nonetheless, it is possible to detect important regularities and to parsimoniously represent the recorded driver data.

Similar to a previous driving simulator study (chapter 3), we employ the distinction between two forms of aberration: violations and errors, as was originally put forward in the context of car driving by Reason et al. (1990). Violations represent deviations from rules that supposedly describe the best/safest way of performing a task (Parker, 2007). Violations do not reflect what the driver *can* do, but what a driver is willing to do; they are – at least in part – intentional. As indicated by Parker (2007): "While many drivers stopped by the police plead that they were unaware that they were breaking the speed limit, it can be argued that at some level we are all aware of the speed we are travelling" (p. 270). Violations are important for safety as they are predictive of a driver's crash involvement (Glendon, 2007; Parker, 2007, for a discussion). A reduction of violations can be best achieved with attention to aspects such as self-control, attitudes, and norms (Parker, 2007).

In contrast, errors are defined as the failure of planned actions to achieve their intended consequences (Reason et al., 1990). They arise from information-processing problems and can be understood in relation to perceptual, attentional, and judgemental processing of the individual (Parker, 2007; Reason et al., 1990). The number of errors can normally be reduced by regular experience or skills training (chapter 3; Parker, 2007).

In addition to violations and errors, the speed of task execution is also an informative driver characteristic. Theories of skill acquisition (see e.g., Crossman, 1959; Groeger & Banks, 2007) predict that beginner performance is slow and errorful. With increasing experience, performance speeds up, and becomes more efficient and less sensitive to distraction. That is, a novice driver will not only make more errors, but will also require more time to complete tasks (e.g., shifting gears; Duncan et al., 1991) than an expert driver. This is in line with the learning curve model that predicts that an operator's required time to complete tasks reduces with increased experience according to a power law function; this model well fit drivers' performance records in a simulator (De Winter et al., 2006b).

Hence, the present study ascribes three generic scores to assess driver proficiency on the basis of performance on the simulator: speed of task execution (or inversely: mean task completion times), violations (i.e., breaking rules, driving too fast), and errors. Not committing violations requires self-control and awareness with respect to the rules; a low number of errors and swift execution of tasks are indicative of handling skill proficiency of the virtual vehicle.

A positive correlation is expected between violations and speed of task execution (chapter 3). That is, those drivers who commit the most violations will tend to be also drivers with low task completion times. Similarly, Reason (1990) found that participants who reported the most violations also tended to rate themselves as particularly skilful drivers. Williams and O'Neill (1974) found that registered racecar drivers (who most likely have excellent vehicle handling abilities and therefore in principle *can* drive very well) were also involved in more violations and accidents than drivers of similar age and gender. This correlation can be explained by the idea that, as novice drivers gain experience, an increase of violations can result (Bjørnskau & Sagberg, 2005; chapter 3), which can be explained by an (over)confidence in one's own skills. An undesired situation arises when individuals use their skills to violate the rules, while failing to recognize hazards that may be present (e.g., Hatakka et al., 2002).

In order to quantify individual violation-scores, it is required that the simulator features a curriculum that allows for self-paced driving. That is, the simulators should *not* simply be counting errors in response to preprogrammed scenarios. Moreover, the simulator measures should be statistically reliable and carry relative validity. It is not essential per se that the simulator replicates the operational environment to a very high accuracy.

1.3. Aim of this study

Our objective was to investigate the statistical relationships between learner drivers' proficiency during simulation-based initial driver training and the on-road driving test, by employing the above framework for modelling the simulator driver. Automatic performance assessments of learner drivers ($N = 804$) were made using the driving simulator – before the students had taken lessons in a real car – and correlated with their performance on the driving test, an average of 6 months later. We expected that success on the driving test is associated in regression analyses with:

(a) a more rapid execution of tasks, (b) a lower number violations, and (c) a lower number of errors: three driver characteristics the combination of which is regarded to be indicative of more proficient simulator driving.

We also investigated statistical associations between simulator proficiency and the duration of on-road training. Previous research has suggested that learner drivers typically apply for their driving test when they consider that they are *just* skilful enough to pass, so that the pass rate depends highly on chance effects (Baughan, 2006). Therefore, we expected that learner drivers with lower simulator proficiency required a longer period to pass their driving test, and that this relationship was stronger than the score on the first driving test itself.

2. Methods

2.1. Background: driver training in the Netherlands

In the Netherlands, training methods to obtain a car driving licence are not subject to stipulations, but accompanied driving without a professional instructor is not permitted. In principle, learner drivers can take their driving test without having any driving lessons, but in practice, nearly everyone takes lessons at commercial driving schools.

Driving tests are organized by the Dutch Driving Test Organization (CBR). The practical part of the test takes 55 min and includes about 35 min of effective driving. A learner driver must pass the theory test before being allowed to take the practical test. Currently, an estimated 100 driving simulators are used to provide driver training in the Netherlands (Kappé & Van Emmerik, 2005). Approximately 3–5% of the learner drivers in the Netherlands have followed part of their training in a driving simulator. Driving simulators are mostly stationed at the larger driving schools. It is legal to take simulator lessons under the age of 18 years, but participating in the CBR theory test as well as taking driving lessons on the road are not permitted before the 18th birthday.

2.2. Driving simulator hardware and software

The present study used medium-fidelity fixed-base simulators of the manufacturer Green Dino (GD) Virtual Realities (Green Dino, 2007). The simulators were stationed at various driving schools all over the Netherlands and were purpose-developed for the initial hours of driver training. The vehicle controls of the simulator resembled those of an actual car with a manual transmission. Force feedback was provided on the steering wheel and acceleration cues were supplied by vibration elements in the steering wheel and the seat. Three projectors provided a horizontal physical 180° field of view. The resolutions were 1024 x 768 pixels for the front view projection and 800 x 600 pixels for the side view projections. The dashboard, interior, and mirrors were integrated in the projected image. The simulation ran at approximately 100 Hz. The screen frame rate was dependent on scene complexity and was high enough to guarantee a smooth dynamic experience.

The curriculum was based on Dutch driver training and consisted of 15 lessons, 27 min each. Lessons 1–5 were dedicated to vehicle control, lessons 6–10 to driving in urban areas with intersections and roundabouts, and lessons 11–15 to motorway driving. Each lesson consisted of two or three so-called blocks, which were preceded by instructive text, voice, and/or movies. During simulation-based driving, a virtual instructor provided (route) instructions as well as feedback on task performance. An important characteristic of the simulator software was that training was not based on preprogrammed scenarios. Students could drive in virtual environments similar to that of a self-paced on-road lesson. The virtual instructor adapted its feedback according to the learner driver's success rate on tasks and the situation in the virtual environment. More information about the virtual instructor, simulator hardware, and the kinds of driving tasks can be found in Weevers et al. (2003b) and in chapter 3.

2.3. Participants

During half-yearly services of the driving simulators at 43 driving schools, copies were made of the driving performance records stored on the computers. We selected the performance records of 2,578 persons who used the same version of the simulator software and whose gender had been filled in on the simulator administration system. On 6 November 2006, the CBR looked up the results of the on-road driving tests and the most recent theory test of the persons in this group whose personal data had also been filled in. Accordingly, the sample used in this study comprised 804 persons (mean age 19.4 years, 54% women) whose simulator performance and CBR test results were available. This study group had their training at 36 driving schools.

Because accompanied on-road practice is not allowed in the Netherlands, and because the simulators were purpose-developed for initial driver training (i.e., before students engage in on-road driver training), it can be assumed that the learner drivers were generally inexperienced with both on-road car driving and simulation-based driver training at the time they started their simulator training.

2.4. Predictor variables

The speed, violation, and error aspects of driver proficiency, which were introduced in section 1.2, were quantified into personal scores using factor analysis on automated simulator performance records. The precise methods for calculating the predictor variables speed-score, violation-score, and steering-error-score can be found in appendix 4A, and were comparable to chapter 3. Only the number of unique blocks that the learner driver had completed was considered; any repeated blocks were not used for calculating the predictor variables. The underlying reasoning was that a drivers' first attempt of a lesson block provided the best indication of his or her actual proficiency.

2.4.1. Speed-score

Speed-score represented a weighted sum of the z-transformed mean task completion times (MTCT) during the simulation-based training programme, with a higher speed-score corresponding to a lower MTCT. In other words, the speed-score represents an aggregate measure of how quickly the trainee executed all tasks in the simulator. Speed-score was calculated using the z-transformed MTCT on the existing task-block-combinations.

2.4.2. Violation-score

Violation-score represented a weighted sum of the z-transformed number of times a learner driver failed in speed-related or safety-related task-block-combinations. Hence, violation-score represented a weighted average of the number of times the learner driver had committed virtual violations.

2.4.3. Steering-error-score

Steering-error-score represented a weighted sum of the z-transformed number of times a learner driver failed in vehicle-control related (steering) task-block-combinations. Note that because the highest factor loadings were found for a few different steering and lane keeping tasks in different training blocks, it was considered most appropriate to refer to this factor as steering-error-score instead of simply ErrorScore.

2.5. Criterion variables

The criterion variables (see Table 1) were the practical driving test results. The present study investigated the relationships with the score on the driving test (ScorePractical) and the on-road training time (LastSim-TestPassed) in most detail. The other criterion variables had more of a descriptive character and were not explicitly associated with the predictor variables.

2.6. Extraneous variables

In addition to the predictor variables, a number of extraneous variables were considered (see Table 2). The extraneous variables were used in regression analyses to minimize the risk of any spurious relationships between the predictor variables and the criterion variables.

Table 1. Criterion variables

ScorePractical [0 = failed, 1 = passed]	This score represents the result of the learner driver's first attempt at the practical driving test on the road.
AttemptsToPass [1, 2, …]	This score represents the number of attempts made to pass the practical driving test.
TestErrors [0, 1, …]	The driving examiner noted the number of errors made during the practical test on a result form (CBR, 2007). The variable TestErrors was defined as the total number of errors made during all the attempts to pass the practical driving test. A learner driver who passed the test had, per definition, 0 errors for that practical driving test.
FirstSim-FirstTest [days]	This variable represents the duration between the first simulator driving lesson and the day of the first practical driving test attempt.
LastSim-FirstTest [days]	This variable represents the duration between the last simulator driving lesson and the day of the first practical driving test attempt.
FirstSim-TestPassed [days]	This variable represents the duration between the first simulator driving lesson and the day that the learner driver passed the practical driving test.
LastSim-TestPassed [days]	This variable represents the duration between the last simulator driving lesson and the day that the learner driver passed the practical driving test.

2.7. Evaluation of data quality of predictor variables

Quality of data associated with the availability of driving test results as well as with missing values and repeated blocks, was evaluated as described below.

2.7.1. Sample bias

Possible issues of sample bias associated with the availability of the driving test results were investigated by comparing the study group of 804 persons and the group (N = 1,774) of simulator drivers whose driving test results were not retrieved, hereafter named the reference group. Differences of the predictor variables and extraneous variables were investigated.

2.7.2. Missing value analysis

Not all learner drivers had completed all simulator blocks, resulting in missing values for the task-block-combinations uncompleted. Most missing data from noncompletion were from simulator blocks later on in the course. Because data of task-block-combinations were z-transformed prior to calculating factor scores, no systematic bias should appear in the predictor variables. To verify this assumption, we

Table 2. Extraneous variables

FirstSimDate [date in days]	For each person, the starting date of the first simulator training was determined. These starting dates were stored automatically on the simulator at the driving schools.
Gender [0 = man, 1 = woman]	An operator entered the gender of each learner driver into the management system of the simulator by hand. The correct gender of the study sample was verified using the CBR data.
Age [days since 18th birthday]	Age was derived by subtracting the learner driver's date of birth from the starting date of the first simulator training (stored automatically). The correct date of birth of the study sample was acquired using the CBR data.
ScoreTheory [44, 45, ..., 50]	Theory tests had to be taken at CBR test locations. A test comprises 50 questions: yes/no, multiple choice, or open questions. The questions were asked orally based on images of traffic situations shown on large TV screens. Candidates could also read the questions on the TV screens. Answers had to be given by means of buttons on the desk in front of the candidate. Results were derived by computer. To pass the test, the candidate had to give a minimum of 44 correct answers. No data were available on (the number of) failed theory tests.
SimBlocks [1 , 2, …, 33]	Driving schools offered simulator lessons in slots of 30 minutes. Each lesson comprised 2 or 3 blocks of 8–15 minutes. The number of unique blocks that the learner driver had completed was recorded. Repeated blocks did not count and were not considered in further analyses.
SimPeriod [days]	The duration of the driver training on the simulator was defined as the starting date of the last simulator lesson minus the starting date of the first simulator lesson. These data were saved automatically on the simulators at the individual driving schools.
PassRateRegion [0–1]	The chance of passing the practical test on the first attempt in the CBR region where the learner driver took the driving test was derived from a website that provided information on the success rates of all driving schools in the Netherlands (CBR, 2008).
TheoryBeforeSim [0 = no, 1 = yes]	This variable described whether the learner driver had passed the theory test before completing the first simulator lesson. The date of the theory test was available from the CBR.

calculated Pearson correlations between the number of unique blocks (SimBlocks) and the predictor variables, for all 2,578 learner drivers of whom simulator records were available. Correlations smaller than 0.10 were considered small enough to assume that there was no systematic bias or statistical artefact, and that these correlations can be adequately explained from, for example, the volunteer bias effects of enrolling in the training programme.

2.8. Statistical analyses

A Pearson correlation matrix was constructed between all the variables to identify significant relations. In order to further verify any possible issues due to missing values, we also included the correlation matrix for a subselection ($N = 296$) of the study group of 804 participants who had completed all lesson blocks and therefore had no missing values (see appendix 4B).

Then, to investigate whether the predictor variables can be used to predict driving test results, robust linear regression analysis was performed on the criterion variables ScorePractical and LastSim-TestPassed. In the linear regression analysis, the predictor variables and the extraneous variables were placed on one side of the equation against one criterion variable on the other side. A weighted sum of the predictor and extraneous variables was calculated to make the most accurate prediction of the criterion variable. The robust linear regression algorithm assigns a lower weight to data points with poor fit; therefore, the results are less sensitive to extreme values. It should be noted that multiple regression makes assumptions about the data, such as normal distribution. These assumptions did not apply to all the variables in this study. However, the quality of the regression analyses is hardly affected when these assumptions are violated (see Gebers & Peck, 2003, for a study with similar methodology). Besides, robust regression analysis is more resistant to violation of the assumptions than regular regression analysis.

Overfitting in the regression analyses was prevented by using the leave-one-out cross-validation (LOOCV) procedure. In this procedure, one observation in the sample was used for validation and the regression analysis was performed on the remaining observations. The procedure was repeated until each observation in the sample had been used once. Then the correlation coefficient was calculated between the predictions of the validation observations and the observed values.

3. Results

3.1. Evaluation of data quality

3.1.1. Sample bias

Table 3 shows the descriptive statistics of the study group ($N = 804$) and the reference group ($N = 1,774$) of simulator trainees whose driving test results were unknown. Gender did not differ significantly between the study group and the reference group ($p = 0.9$). However, the study group completed significantly more training blocks (SimBlocks) on the simulator than the reference group ($p < 0.001$), whereas the duration of training (SimPeriod) showed no significant difference ($p = 0.2$). Moreo-

Table 3. Descriptive statistics on the study group and the reference group whose driving test results were not retrieved

	Study sample (*N* = 804)				Reference (*N* = 1,774)	
	Mean	SD	Min	Max	Mean	SD
Predictor variables						
Speed-score	0.118	0.905	-6.03	2.64	-0.053	1.04
Violation-score	0.010	0.960	-2.35	5.33	-0.005	1.02
Steering-error-score	-0.184	0.972	-2.43	4.38	0.083	1.00
Extraneous variables						
FirstSimDate	1-Nov-05	54	1-Aug-05	6-Mar-06	7-Nov-05	58
Gender	0.545	0.498	0	1	0.548	0.498
Age[a]	505	1240	-227	9610	–	–
ScoreTheory[b]	46.4	1.75	44	50	–	–
SimBlocks	23.4	10.9	1	33	21.7	12.6
SimPeriod	19.5	19.0	0	135	21.0	30.7
PassRateRegion[b]	0.491	0.051	37.5	60.0	–	–
TheoryBeforeSim[b]	0.131	0.337	0	1	–	–
Criterion variables						
ScorePractical[b]	0.540	0.499	0	1	0.529[d]	–
AttemptsToPass[b,c]	1.65	0.984	1	6	–	–
TestErrors[b]	4.47	6.96	0	50	–	–
FirstSim-FirstTest[b]	166	84	12	420	–	–
LastSim-FirstTest[b]	146	85	10	412	–	–
FirstSim-TestPassed[b,c]	185	90	14	455	–	–
LastSim-TestPassed[b,c]	166	90	10	447	–	–

[a] Days since 18th birthday

[b] These data were not available for the reference group; the reference group comprises learner drivers whose test results were unknown.

[c] In 77 persons, the number of days and the number of attempts until they passed their practical test were not determined, because they had not yet passed the practical test. These learner drivers may have still been having driving lessons, or they may have quit.

[d] This number was estimated from a website that contains the success rates of all the driving schools in the Netherlands (CBR, 2008). We selected those driving schools with a GD simulator. Note that in 2005, the national average chance of passing the practical test first time was 48.6% (CBR, 2005).

ver, the study group had a slightly higher speed-score and a lower steering-error-score than the reference group ($p < 0.001$). Possible explanations for these differences are that the study group, per definition, comprised learner drivers who took

Table 4. Correlations between number of simulator blocks and the predictor variables (N = 2,578)

	Speed-score	Violation-score	Steering-error-score
SimBlocks	0.001 (p = 0.9)	-0.059 (p = 0.003)	-0.004 (p = 0.9)

the theory test and practical test within 1 year, and whose personal data had been filled in fully and accurately in the driving school's administration system. Therefore, it is possible that the study group comprised persons who took the driving simulator training more seriously than the reference group, whereas the latter group could be contaminated with invalid test lessons, demonstration lessons, and so forth.

The study group's mean score on the first attempt to pass the practical driving test (ScorePractical) was 0.540 (54%). Verification on the basis of a website that contains the success rates of all the driving schools (CBR, 2008) showed that learner drivers who took their driving test via a driving school with a GD simulator had a 52.9% chance of passing in the period April 2006 – March 2007, a statistically insignificant difference with regard to the study group (95% confidence interval of ScorePractical was 0.505 to 0.575, assuming binomial distribution).

It can be concluded that the study group did not deviate from the reference group for gender, duration of simulation-based training, and pass-rate on the driving test. Differences in driving performance between the study group and the reference group were considered small enough to assume that *relative relations within* the sample were representative for all GD simulator trainees.

3.1.2. Missing value analysis

Table 4 shows the Pearson correlations between the predictor variables and the number of blocks. The correlations were considered small enough to assume that there was no systematic bias in the predictor variables as a consequence of the missing values.

3.2. Correlations between variables

Table 5 shows the matrix of correlations between the variables defined in sections 2.4 to 2.6. Noteworthy correlations are described in brief below.

A higher pass rate (ScorePractical) significantly correlated with a higher speed-score and a lower steering-error-score. Correlations between the predictor variables and the duration of driver training were generally stronger than the correlations with ScorePractical. Passing the test in a shorter time (lower LastSim-TestPassed) corresponded with a higher speed-score, a lower steering-error-score, but with a higher violation-score.

Note that violation-score and speed-score were strongly confounded (r = 0.68, p < 0.001). The correlation between violation-score and LastSim-TestPassed was

Table 5. Pearson correlations (*N* = 804) (multiplied by 100 for convenience)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1. FirstSimDate																	
2. Gender	-3																
3. Age	-12*	1															
4. ScoreTheory	5	-6	5														
5. SimBlocks	5	5	-4	6													
6. SimPeriod	6	5	-3	3	44*												
7. PassRateRegion	4	0	-3	0	0	10*											
8. TheoryBeforeSim	1	-5	6	2	2	18*	9*										
9. Speed-score	14*	-44*	-18*	2	9*	5	2	5									
10. Violation-score	5	-41*	-10*	-1	-2	4	-3	5	68*								
11. Steering-error-score	-9*	33*	6	-12*	-13*	-7*	-4	-2	-28*	-4							
12. ScorePractical	6	-2	-2	11*	6	-1	6	2	8*	-2	-19*						
13. AttemptsToPass	-9*	3	1	-8*	-9*	-3	-1	3	-9*	1	17*	-80*					
14. TestErrors	-8*	1	12*	-9*	-10*	-4	-4	0	-12*	2	24*	-70*	95*				
15. FirstSim-FirstTest	-15*	11*	9*	-5	-11*	9*	-22*	-18*	-22*	-4	20*	-5	-8*	3			
16. LastSim-FirstTest	-16*	10*	10*	-6	-21*	-13*	-25*	-22*	-23*	-5	21*	-5	-7*	4	97*		
17. FirstSim-TestPassed	-19*	11*	14*	-4	-13*	7	-18*	-14*	-27*	-7	24*	-32*	39*	40*	85*	83*	
18. LastSim-TestPassed	-21*	10*	15*	-5	-22*	-14*	-21*	-17*	-28*	-8*	26*	-31*	39*	40*	83*	86*	98*

* $p < 0.05$

actually positive when partialling out the speed-score ($r = 0.16$, $p < 0.001$). Hence, these results support the proposed theoretical framework indicating that violations are a negative trait. Note that the steering-error-score and speed-score had a moderate negative correlation ($r = -0.28$, $p < 0.001$), which is in accordance with results in mental chronometry. Although speed and accuracy are well known as a negative correlation *within* individuals (i.e., the so-called speed-accuracy trade-off), a positive correlation has been found *between* individuals (i.e., the faster persons are generally also those persons who make less errors, given equal tasks and instructions) (Jensen, 2006).

A later date of the first driving simulator lesson (i.e., later in the study period) negatively correlated with the time until the candidates took their practical test (last four criterion variables). This could be explained by the measurement set-up. The study group completed their first simulator lesson between August 2005 and March 2006, whereas the driving test result data were collected on November 2006. For persons who had simulator lessons at an earlier date, there was a greater chance that their practical test result would be available. The relations observed with the date emphasize the importance of controlling for this variable in regression analyses.

There were wide differences in performance between men and women. Women had a lower speed-score, a lower violation-score, and a higher steering-error-score. Learner drivers who completed a larger number of simulator training blocks passed their driving test sooner, irrespective of whether we took the date of their first simulator training into consideration, or their last.

An interesting finding was that a longer duration of driver training to the first test (FirstSim-FirstTest and LastSim-FirstTest) was not correlated with the score on the first practical driving test (ScorePractical). This can probably be explained by the fact that learner drivers continued to train until their level of expertise was high enough to pass their first practical driving test (e.g., Baughan, 2006).

Appendix 4B shows the correlations for the learner drivers ($N = 296$) from the study group ($N = 804$) who had completed all 33 simulator blocks. It can be seen that the patterns of correlations were roughly identical to Table 5. Correlations with the predictor variables tended to be somewhat stronger, which can be adequately explained considering that the predictor variables could be estimated more reliably when data did not have missing values.

3.3. Regression analyses for the prediction of criterion variables

The correlation matrix showed relations between performance on the simulator and the driving test results. A subsequent regression analysis was carried out to predict ScorePractical and LastSim-TestPassed. By conducting a regression analysis, we controlled for extraneous variables, such as FirstSimDate.

Table 6. Regression analysis on criterion variable ScorePractical (N = 804)

Variable	Coefficient	Standard error	t-score	p-value
Speed-score	0.0504	0.0311	1.62	0.106
Violation-score	-0.0315	0.0282	-1.12	0.264
Steering-error-score	-0.0941	0.0218	-4.32	<0.001
FirstSimDate	0.000363	0.000353	1.03	0.304
Gender	0.0623	0.0449	1.39	0.166
Age	-1.56×10^{-6}	1.55×10^{-5}	-0.10	0.920
ScoreTheory	0.0258	0.0108	2.38	0.017
SimBlocks	0.00233	0.00196	1.19	0.235
SimPeriod	-0.00179	0.00113	-1.58	0.113
PassRateRegion	0.576	0.374	1.54	0.124
TheoryBeforeSim	0.0439	0.0570	0.77	0.442

Note. Correlation between predicted and observed values of ScorePractical = 0.239 (p < 0.001). Leave-one-out cross validation correlation = 0.184 (p < 0.001) (excluding predictor variable FirstSimDate).

3.3.1. Regression analysis on ScorePractical

Table 6 shows the results of the regression analysis to predict ScorePractical. A higher success rate was found to be significantly associated with a lower steering-error-score on the simulator and a higher score on the theory test.

To verify the predictive strength of the regression model, the analysis was repeated after exclusion of FirstSimDate, because it showed masked dependence on the criterion variables. The leave-one-out cross-validation (LOOCV) procedure was used to test the strength of the prediction. The LOOCV correlation was 0.184. Table 7 shows the cross-tabulation of the predictions obtained using the cross-validation procedure. To maximize the total prediction, we chose a classification threshold that distinguished between the candidates who passed and the candidates who failed. In this way, the correct driving test result could be predicted in 60.8% of the candi-

Table 7. Cross-tabulation of predictions and observed values of ScorePractical (N = 804)

	Predicted				
Observed	Failed	Passed	Row total	Correct predictions as % of row total	Correct predictions as % of grand total
Failed	170	200	370	45.9%	
Passed	115	319	434	73.5%	
Column total	285	519			60.8%

Note. Chi-square = 33.0 (p < 0.001)

Table 8. Regression analysis on criterion variable LastSim-TestPassed (*N* = 727)

Variable	Coefficient	Standard error	*t*-score	*p*-value
Speed-score	-26.8	5.05	-5.31	<0.001
Violation-score	10.4	5.50	2.32	0.021
Steering-error-score	15.6	3.65	4.27	<0.001
FirstSimDate	-0.196	0.0571	-3.44	<0.001
Gender	-6.33	7.29	-0.87	0.386
Age	0.00817	0.00273	2.99	0.003
ScoreTheory	-0.270	1.75	-0.15	0.878
SimBlocks	-1.41	0.317	-4.45	<0.001
SimPeriod	0.012	0.181	0.64	0.523
PassRateRegion	-331	60.8	-5.44	<0.001
TheoryBeforeSim	-43.6	9.11	-4.78	<0.001

Note. Correlation between predicted and observed values of LastSim-TestPassed = 0.491 (p < 0.001). Leave-one-out cross validation correlation = 0.448 (p < 0.001) (excluding predictor variable FirstSimDate).

dates. Another method to choose a threshold is by equalising the number of Type II errors. This led to correct predictions in 58.6% of the candidates. Note that the study group's average score on the first practical driving test was 54.0% (see Table 3).

3.3.2. Regression analysis on LastSim-TestPassed

Regression analysis was also performed on LastSim-TestPassed. The results are shown in Table 8. Earlier success in passing the practical part of the driving test, measured from the last simulator lesson was associated with a higher speed-score, a lower violation-score, a lower steering-error-score, a later date (later FirstSimDate, i.e., later in the study period), younger age, completing more training blocks on the simulator, taking the driving test in a region with a higher chance of success, and passing the theory test before the last simulator lesson. As was also the case with ScorePractical in section 3.3.1, the regression analysis was repeated after excluding FirstSimDate. The LOOCV correlation was 0.45 (p < 0.001).

4. Discussion

Significant quantitative relationships have been found between performance on the simulator and the driving test results. Fewer steering errors corresponded to a higher chance of passing on the first attempt of the driving test. Fewer steering errors, fewer violations, and faster task execution corresponded to a shorter duration of driver training. Hence, the present findings support the theoretical framework from section 1.2 that expressed increased driver proficiency in terms of higher speed of task execution, fewer violations, and fewer errors.

4.1. Size of the predictions

The predictive strength of the individual variables analysed in this study was moderate, with correlations up to 0.28 between individual predictor variables and the criterion variables. Lew et al. (2005) found higher correlations in the range of 0.6 to 0.8 between performance on the simulator and human assessment after 10 months. Our study formed a heavier test case, however, because the study by Lew et al. targeted a rather heterogeneous group of patients with moderate to severe traumatic brain injury. In our case, the study group was more homogeneous: the learner drivers were relatively healthy, were of similar age, and all had the same goal of passing their driving test. Owing to range restriction effects, individual differences could have been less prominent.

Furthermore, it is the intention that learner drivers adapt their abilities and attitudes during driver training. The driving simulator in the present study was used *before* the learner drivers had even taken a lesson in a real car. Therefore, associations between performance on the simulator and the driving test result may have been confounded.

In addition, it was not necessarily the predictor variables (i.e., the simulator measures) that were unreliable, but more likely, it was the criterion variables instead. A study in the UK (described in Baughan, 2006) required that 366 individuals retake the practical part of the driving test within a few days of the first test, with a different examiner, but in the same region. The result of the first driving test was not revealed until after the retest. During the retest, the examiner was not aware of the result of the first test. The results showed that the consistency of the practical driving test was 64%; 16% passed the first test and failed the second; 20% passed the second test, but failed the first. On the basis of the chance of 37.4% of passing the first test and of 42.3% of passing the second test, a consistency of 51.9% would be expected if totally random data were used. A partial explanation given for the strikingly low consistency of the practical driving test was that learner drivers apply to take their test when they consider their level of expertise to be just sufficient to pass (Baughan et al., 2005; Baughan, 2006). The test result therefore depended strongly on coincidental circumstances.

Reliability of the criterion variables may be even further reduced by regional differences. Baughan et al. (2005) found that there were differences in the duration of the driving test, the items addressed during the test, and the pass rates between test centres at different locations. It is unclear to what extent these values are representative of the Dutch situation. Nonetheless, it can be considered that our 60.8% correct prediction rate of the driving test score is realistic and relatively strong.

Tentatively, when considering the correlation of -0.23 between steering-error-score and ScorePractical in appendix 4B and when assuming that the steering-error-score was found to be reliable, applying a correction due to unreliability of the criterion variable (test-retest reliability 0.25 based on Baughan, 2006) yielded a more promising correlation coefficient of -0.47. Improving the reliability of the driving test could be a fruitful effort for improving its worth as a criterion.

The duration of driver training was much easier to predict than the result of the practical driving test. The cross-validation correlation coefficient of 0.45 implied that 20% of the variance could be explained by linear regression. The observation that the duration of driver training was easier to predict than the test result can also be explained on the basis of the learner drivers having barely sufficient expertise to pass the test.

4.2. Validity of the on-road driving test

We employed learner drivers' simulator measurements to predict the outcome on the driving test. A very important question that remains, however, is whether the test score validly discriminates between safe and unsafe drivers. A literature overview provided by Senserrick and Haworth (2005) showed that there is generally little association between scores of on-road assessments and crash rates once licensed. The disconnect between test scores and road safety is evident when studying gender differences: although young men have a higher crash risk than young women, they tend to have a *higher* pass rate on the driving test (Crinson & Grayson, 2005).

Yet, on the other hand, a meta-analysis of Elvik and Vaa (2004) found that the more stringent tests could discriminate well between safe and unsafe drivers. Moreover, Baughan and Sexton (2001) found a strong predictive relationship between faults on the driving test and crash risk once the effects of age, mileage, and driving in the dark (all being risk factors) were statistically adjusted for. In conclusion, the relation between test success and crash risk appears to be rather complex and dependent on several personal factors.

Instead of predicting the score on the driving test, it could be more informative to directly investigate the statistical relationship between performance in the simulator and crash involvement during later driving. Collecting reliable and unbiased accident data, however, is a difficult undertaking, which, paradoxically, has led researchers to recommend using alternative measures for identifying accident-prone drivers, such as closed-course performance or driving simulator performance (Ranney, 1994). Another perhaps more promising possibility is to investigate the association between simulator performance and unobtrusively recorded on-road performance (see e.g., Yan et al., 2008). It is recommended to increase research effort into investigating the validity of driving simulators and on-road driving tests, with a view towards improving road safety.

4.3. Driver performance and driver behaviour

The present study ascribed three generic scores on the basis of simulator training results. One may argue that this is an oversimplified representation for a task as complex as car driving. According to Groeger (2000a), driving a car is a complex task requiring many different skills. He found, for example, that persons who learn to negotiate crossroads well do not automatically learn to also negotiate roundabouts well. These results are supported in the present study. Each task has a substantial level of uniqueness. It is a question of how generic/specific the researchers wish to be in their assessments.

There were several reasons why we used a simple framework in this study. Overviews have shown that the post hoc character of some studies have contributed to the lack of progress in the development of theories on driving behaviour (e.g., Ranney, 1994). When the level of complexity of a model is ignored, there is a risk of overfitting and limited generalizability (Preacher, 2003). In addition, the simulator measures used in this study were theoretically plausible. Speed and accuracy (inversely: errors) are two measures related to information-processing capacity, which can be optimized accordingly by a human operator (Fitts, 1954; Jagacinski & Flach, 2003; Salthouse, 1979). De Winter et al. (2006d) and Zhai et al. (2004) showed that the same principle applied to steering on a winding road on a driving simulator. Recently, the predictive validity of measures that represent information-processing capacity has been (re)recognized. A meta-analysis demonstrated that the Useful Field Of View test (UFOV) was a strong predictor of driving behaviour and crash involvement amongst older drivers (Clay et al., 2005). It is therefore a logical choice to incorporate an individual's information-processing capacity into driver models.

It is acknowledged that vastly different factors govern the crash risk of young drivers. Young drivers, and particularly young male drivers, have greater tendencies towards risk factors such as sensation seeking, speeding, and being influenced by peers (Hatakka et al., 2002; J.D. Lee, 2007; OECD, 2006).

It has been argued that such behavioural aspects determine safety, and that simulators and on-road driving tests tend to measure performance in terms of skill-based control abilities and so may be poor indicators of safety (Evans, 2004; Senserrick & Haworth, 2005). As was noted by Evans (2004): "Simulators measure driver performance, what the driver can do. However, safety is determined primarily by driver behavior, what the driver in fact chooses to do. It is exceedingly unlikely that a driver simulator can provide useful information on a driver's tendency to speed, drive while intoxicated, run red lights, pay attention to non-driving distractions, or not fasten a safety belt. Twenty-year-olds perform nearly all tasks on simulators better than the 50-year-olds, but it is the 50-year-olds who have sharply lower crash risks" (p. 188).

As mentioned in the introduction, we agree that driving simulators measure driver performance in terms of information-processing capacity. However, we disagree that a simulator can measure *only* driver performance. In our study, as shown in Table 5, younger drivers failed significantly *more* often in the violation tasks. This supports that violations indeed represented a behavioural aspect of driving, in correspondence with earlier research (chapter 3).

Another relevant observation was that the simulator distinguished strongly between men and women. Men showed faster task execution and committed more violations but made fewer steering errors. This is in agreement with earlier research into driving tuition in the UK (Maycock & Forsyth, 1997), where women were found to make more vehicle-control errors during their driving test than men. It is well known that, on the road, men commit more violations and are involved in more speed-related (fatal) crashes than women (OECD, 2006), whereas women show a greater tendency towards involvement in accidents while manoeuvring at low speed (Kim, 1996; Laapotti & Keskinen, 2004). The gender differences on the simulator were clearly interpretable as far as this was concerned. It should be noted that the present gender differences were probably mediated by differences in prior experience in human-computer interaction or with prior on-road experience in (e.g., mopeds) as well.

Finally, the driving simulator provided information on whether the learner driver had optimized accuracy or whether he or she optimized speed in the driving task. The distinction between accuracy (errors) and speed/violations may partly overlap with other taxonomies in traffic psychology, such as driver performance and driver behaviour (Evans, 2004), driving skill and style (Elander et al., 1993), skills and safety motives (Lajunen & Summala, 1995), and the notion that driving is not only about what a driver can do, but also about what a driver is willing to do (Ranney, 1994).

It is concluded that it is possible to identify individual differences in driver performance *and* driver behaviour during simulator training, even before a person has driven a real car and that this may be highly relevant within the framework of road safety.

4.4. Driving simulator fidelity and data quality

An often heard critique is that specifications of training simulators are too much technology-driven, whereas functional specifications remain vague (e.g., Verstegen, 2003). The present application of a statistical approach goes beyond engineering-oriented approaches in simulator requirements. The use of factor analysis supplied individual scores of driving proficiency that were predictive of driving test results; hence, providing relevant information regarding the functional use of simulators for driver training and testing. In essence, the value of a simulator is determined by the quality of learning and the quality data it produces, not by physical fidelity per se.

This study was performed using a relatively low-cost simulator. Earlier research into the validity of simulators also made use of low-cost simulators (e.g., H.C. Lee, 2003; Lew et al., 2005). Why is it that low-cost driving simulators seem to be successful in driver assessment? High-fidelity simulators offer redundant information such as motion cues and complex visual scenes. Beyond the fact that high-fidelity is usually associated with high financial cost and that the value of full motion cueing for training is still debated, the question arises as to whether higher fidelity simulators always yield higher quality data (even though being more realistic). Contrary, psychometric measurements generally take place according to highly standardised procedures, in which the matter is to exclude as many irrelevant influences and as much randomness as possible. This can be placed within the framework of one of the statements made by J.D. Lee (2004). He considered that high-fidelity simulator training can become diluted and can make it more difficult to gather data. Summarized, copying reality as closely as possible is not always better.

4.5. Study limitations

In the present study, we did not control the driver training on the road. Events that took place in this intervening period were unknown, such as the number of driving lessons on the road, the type of driving lessons, what type of cars the candidates drove, and which human instructors supervised the training programme. The lack of objective observations and standardization during road training is an inherent characteristic of on-road driver training. With the aid of regression analysis, we controlled for extraneous variables, such as date, age, gender, and the number of simulator lessons completed by the learner drivers.

In this study, it was found that the sample of simulator trainees had a 4–5% higher chance of passing their driving test than the national average. Moreover, regression analysis showed a relatively strong negative association between the number of simulator training blocks and the duration of driver training. These observations do not prove that the use of a simulator causes better driving test results. It is also possible that the better driving schools were more likely to have a simulator, and that persons with superior ability were more inclined to make use of the driving simulator. The present sample of students was relatively small compared to the 170,000 candidates who succeed on the practical driving licence test each year in the Netherlands (CBR, 2005). Volunteer bias may have played a role in the decision to drive in the simulator. Therefore, generalization of the present results to the entire young driver population should be done with careful consideration. One particular characteristic of the study group was that many of these students started the simulator training before their 18th birthday, whereas on-road driver training is not allowed in this situation.

It has to be stressed here that the aim of this study was not to demonstrate causality, but to evaluate the statistical associations between simulator performance and driving test results. The fact that relations were found despite the presence of many uncontrolled factors actually strengthens the confidence in the present results.

4.6. Recommendations

The results indicate that simulator-based performance measurements on a learner driver's actions can provide measures that are informative of future on-road driving test performance. These measures could potentially be applied for summative or formative assessments of the learner driver in order to improve driver training effectiveness. A speed-score, violation-score, and steering error-score may be regularly outputted by the driving simulator and used for feedback on performance. It is yet to be investigated how such measures can be used most effectively in practice. One interesting possibility is to provide a special training programme to those students with a high violation-score. Here, one may think of training of self-control (Hatakka et al., 2002), a driver coaching system to change attitudes (Stanton et al., 2007), or a training programme that provides insights into potential risks and deflates overconfidence (Senserrick, 2001).

Driving is an important and safety-critical task, with young and inexperienced drivers being overly involved in crashes (e.g., Mayhew et al., 2003; Williams, 2003). Remarkably, diverse overviews and a meta-analysis showed that taking driving lessons on the road from a professional instructor did not result in fewer crashes compared to driving under the supervision of a nonprofessional (Brown, 1997; Elvik & Vaa, 2004; Evans, 2004). One of the disadvantages of driver training is that there is poor perception of what actually happens on the road and a lack or absence of objective measurements. There is sufficient rationale why driving simulators should be taken seriously as complementary training and assessment tools. It is recommended that researchers conduct more large-scale validation studies.

Appendix 4A

This appendix describes how we calculated the speed-score, violation-score, and steering-error-score.

Speed-score

The instruction software had 209 driving tasks and 33 training blocks. For each person and for each existing task-block-combination, the MTCT of correctly performed tasks was calculated. Trivial variables with a standard deviation of the MTCT of less than 1 s were excluded. Next, redundant variables that had strong correlations (> 0.8) were excluded in order to obtain a more valid and diverse set of variables (inspired from Boyle, 1991), leaving 457 variables. The resulting data matrix (2,578 persons x 457 task-block-combinations) was z-transformed and the pairwise Pearson correlation matrix was submitted to principal axis factoring from which one factor was extracted, which explained 9.1% of the variance.[1] The decision to extract one factor was supported by the reduced scree plot and the good interpretability of the loadings compared to those obtained after excluding more than one factors.[2]

33 of 457 factor loadings were higher than 0.5, which were self-paced tasks in which there were wide individual differences in MTCT. These tasks included lateral motorway manoeuvres (i.e., filtering in, filtering out on motorways, changing lanes), changing gear, driving along straight stretches of road in urban or rural conditions, and negotiating intersections. These results were qualitatively congruent with earlier research using another sample and fewer variables for calculating a speed-score from MTCTs (chapter 3). A speed-score was calculated using the Bartlett procedure, based on the z-transformed data matrix.

Violation-score and steering-error-score

First, we selected the number of failed tasks on all the task-block-combinations. The easy and rare tasks were removed by excluding those task-block-combinations with a standard deviation of the number of failures smaller than 0.5. Next we excluded

[1]) The solutions explained a rather small share of the total variance. It must be stressed, however, that the aim of factor analysis is not to extract as much variance as possible, but to identify latent patterns. The low variances were caused by the fact that we included many variables featuring relatively low communalities. The decision to include many variables was based on Monte Carlo simulations from which it was found that factor scores could be most reliably approximated by submitting as much information possible to the factor analysis. Low communalities were partially caused by the fact that the data were obtained from events occurring relatively rarely in a task-block-combination, thereby having low statistical power. Communalities could easily have been raised by excluding low-communality variables, and/or by parcelling. Further discussion on this topic was provided in chapter 3.

[2]) The eigenvalue of the first factor was 150. The eigenvalue of the second, unretained factor was 48.2.

variables with strong correlations (> 0.8), leaving 410 variables. The resulting 2,578 x 410 data matrix was z-transformed and subjected to principal axis factor analysis with oblique direct quartimin rotation to extract two factors. The decision to extract two factors was supported by the reduced scree plot and the interpretability of the loadings as compared to the one or three factor solutions.[3]

As the loadings featured a small correlation (-0.072), the solution was simplified using orthogonal Varimax rotation. The first factor explained 4.9% of the variance. 32 of the loadings were higher than 0.4. These comprised exceeding the speed limit, driving too close to the vehicle in front, and letting the clutch out too quickly, particularly during motorway training blocks. Hence, the first factor could be interpreted as a violation factor in which speed or virtual safety margins had been exceeded, in qualitative agreement with earlier research (chapter 3). A violation-score was calculated for each learner driver using the Bartlett procedure.

The second factor explained 3.0% of the variance and 23 loadings were higher than 0.4. These comprised too much deviation from the centre of the road in bends and on straight stretches or tortuous roads in urban or rural conditions. The second factor could be interpreted as a steering error factor related to elementary vehicle control. A steering-error-score were calculated for each learner driver using the Bartlett procedure.

Figure A1 shows the factor loading plot of the violation and steering-error factors.

Reliability and invariance of factor analyses

Reliability of the factor analysis solutions (speed-score, violation-score, steering-error-score) was evaluated by dividing the samples at random into two equal halves and calculating factor scores on the basis of the loadings on the two subsamples. Then comparability coefficients were calculated, that is, the correlation between the factor scores (Everett, 1983). This procedure was repeated 10 times. Not all the learner drivers had completed all simulator blocks, so the data contained missing values. Therefore, additionally, comparability coefficients were calculated between the candidates who had completed all blocks in the standard order and the candidates who had not. Furthermore, comparability coefficients were calculated between men and women, because previous research has shown large gender differences in the speed-score, violation-score, and steering-error-score (chapter 3). Finally, comparability coefficients were calculated between the learner drivers in the study group and the group of simulator learner drivers whose test results could not be determined (reference group). In all cases, a comparability coefficient of greater than 0.90 was considered to provide sufficient evidence of factorial invariance of the subsamples.

[3]) The eigenvalues of the first and second factor were 20.6 and 12.8, respectively; the eigenvalue of the third, unretained factor was 7.1.

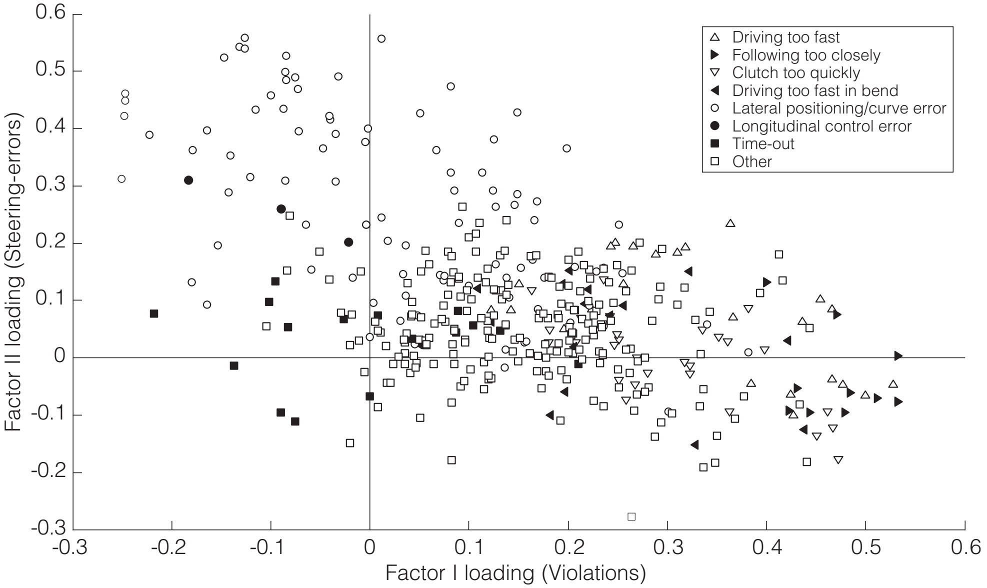


Figure A1. Factor loading plot of Violations and SteeringErrors (410 task-block-combinations).

Results of the analyses are shown in Table A1. Comparability coefficients were high enough to conclude that the factor scores were reliable and invariant in the subsamples.

Table A1. Reliability and factorial invariance of subsamples based on comparability coefficients

	Random split half [a] *N* = 1,289/1,289	Default/ non-default *N* = 481/2,097	Men/women *N* = 1,167/1,411	Sample/ reference *N* = 804/1,774
Speed-score	0.997	0.994	0.994	0.984
Violation-score	0.991	0.986	0.982	0.994
Steering-error-score	0.983	0.984	0.977	0.985

[a] Mean of 10 repetitions

Appendix 4B

Table B1.

Table B1. Pearson correlations (multiplied by 100 for convenience) for those learner drivers from the study group who completed all 33 simulator blocks ($N = 296$)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1. FirstSimDate																	
2. Gender	-7																
3. Age	-20*	9															
4. ScoreTheory	-1	-9	2														
5. SimBlocks [a]																	
6. SimPeriod	5	6	-2	-4													
7. PassRateRegion	10	7	0	10		1											
8. TheoryBeforeSim	10	-12*	9	0		16*	1										
9. Speed-score	20*	-52*	-22*	5		-1	-8	5									
10. Violation-score	6	-47*	-12*	6		-1	-7	2	70*								
11. Steering-error-score	-6	44*	3	-18*		2	-2	-4	-38*	-11*							
12. ScorePractical	9	-9	-7	18*		-1	14*	5	13*	4	-23*						
13. AttemptsToPass [b]	-12*	7	1	-15*		-3	-15*	3	-17*	-4	18*	-82*					
14. TestErrors	-9	6	13*	-8		0	-13*	2	-21*	1	26*	-69*	95*				
15. FirstSim-FirstTest	-14*	15*	6	1		31*	-19*	-9	-16*	3	24*	-3	-6	5			
16. LastSim-FirstTest	-16*	14*	6	1		8	-20*	-14*	-16*	4	24*	-3	-5	6	97*		
17. FirstSim-TestPassed [b]	-20*	16*	15*	-2		22*	-21*	-5	-23*	-2	29*	-33*	40*	42*	85*	84*	
18. LastSim-TestPassed [b]	-23*	15*	18*	-2		0	-22*	-9	-24*	-2	29*	-34*	42*	43*	81*	85*	98*

* $p < 0.05$

[a] Correlations with SimBlocks could not be computed because SimBlocks was equal to 33 for the entire subselection.

[b] In 23 persons, the number of days and the number of attempts until they passed their practical test could not be determined, because they had not yet passed the practical part of the test. These learner drivers may have still been having driving lessons, or they may have quit.

CHAPTER 5

Advancing simulation-based driver training: Lessons learned and future perspectives

Abstract

This chapter aims to provide recommendations for improving the effectiveness of automatic, student-adaptive, simulation-based driver training. Using experiments and recorded data in driving simulators, three distinct issues are evaluated: (a) the student, (b) the virtual driving instructor (VDI), and (c) the student-profile. We found that: first, students seek task-relevant information themselves; they can learn some tasks without additional feedback and instructions. Second, an intelligent VDI that emulates a human driving instructor is not favoured. To the contrary, regressive instruction – a relatively simple principle – was effective in letting students drive away autonomously. Third, constructing a student-profile based on individual characteristics, such as a strength-weakness report, is viable for providing student-adaptive feedback.

De Winter, J.C.F., De Groot, S., Dankelman, J., Wieringa, P.A., Van Paassen, M.M., & Mulder, M. (2008a). Advancing simulation-based driver training: lessons learned and future perspectives. *Proceedings of the 10th International Conference on Human-Computer Interaction with Mobile Devices and Services* (*Mobile HCI 2008*), Amsterdam, the Netherlands, 459–464. (adapted with minor textual changes)

1. Introduction

The greater part of current driver training takes place on the road under the supervision of a human instructor. This traditional form of training is expensive, whereas research has shown that it does not reduce post-licence crash risk as compared to informal training (Elvik & Vaa, 2004). Driving simulators are a complementary tool to on-road training and offer advantages such as objective student assessment, standardization, free control over the training conditions, potential cost-effectiveness due to automation, and didactic possibilities such as multimodal feedback, demonstrations, and replays (chapter 1; De Groot et al., 2007). However, research and development of simulators is primarily hardware-oriented (e.g., developing display techniques), whereas the aforementioned advantages are not optimally exploited (chapter 1). It is argued that research is needed to find out how to get the most out of simulators so that students are efficiently trained to obtain their licence and to drive safely.

At present, the Netherlands has an important role in the domain of simulation-based driver training (Kappé & Van Emmerik, 2005). A major player is Green Dino Virtual Realities. Their Dutch Driving Simulator (DDS; Green Dino, 2007) is mainly used for automatic training at driving schools across the country. In this training mode, a virtual driving instructor (VDI) provides feedback and instructions during the training sessions (Weevers et al., 2003b). A human supervisor has the possibility to evaluate a student's progress after a simulator-session has been completed and accordingly can decide to alter the training curriculum, to repeat a lesson, or to transfer to a real car.

This chapter aims to provide recommendations about how to improve the effectiveness of automatic, student-adaptive, simulation-based driver training systems. The focus of this chapter is not on technical hardware requirements, but on didactic software requirements instead. Using experiments and recorded data in driving simulators, three distinct issues are evaluated:

1. *The student*. Considering that humans are intelligent systems, it is first explored what students can do by themselves, and when they actually need feedback and instructions.
2. *The virtual driving instructor (VDI)*. It is investigated how the VDI can be improved. Should the VDI's intelligence be increased so that it better understands the human student? The VDI's effectiveness is studied in a specific task (driving away) and suggestions of improvement are provided.
3. *Constructing a student-profile*. Herein, we investigate whether a student-profile based on individual differences in driver behaviour can be used for advancing training effectiveness.

Even though a simulator is a stationary system (providing an illusion of being on the move), the results in the present study bear direct relevance to human-computer interaction with mobile devices. A student-profile is portable and can possibly be used in advanced driver assistance systems (ADAS) or future automotive navigation systems.

2. The student

2.1. Self-paced task

Humans are active information seekers rather than passive recipients. This presumption also applies to simulation-based driver training, because car driving is, to a large extent, self-paced (Fuller, 2005). Previous research showed that there were considerable differences between students in driving simulation performance, with respect not only to success rates on driving tasks, but also to driving speed and speed of task execution (chapter 3). To illustrate, Figure 1 shows students' (*N* = 1,760) time to complete a lap around a square block of intersections during a *turn right* exercise using computerized driving instruction. There was no other traffic. The data have been obtained by DDSs stationed at driving schools in the Netherlands in the period August 2005 – March 2006. It can be seen that there were considerable individual differences in driving speed, indicating that students were partially re-

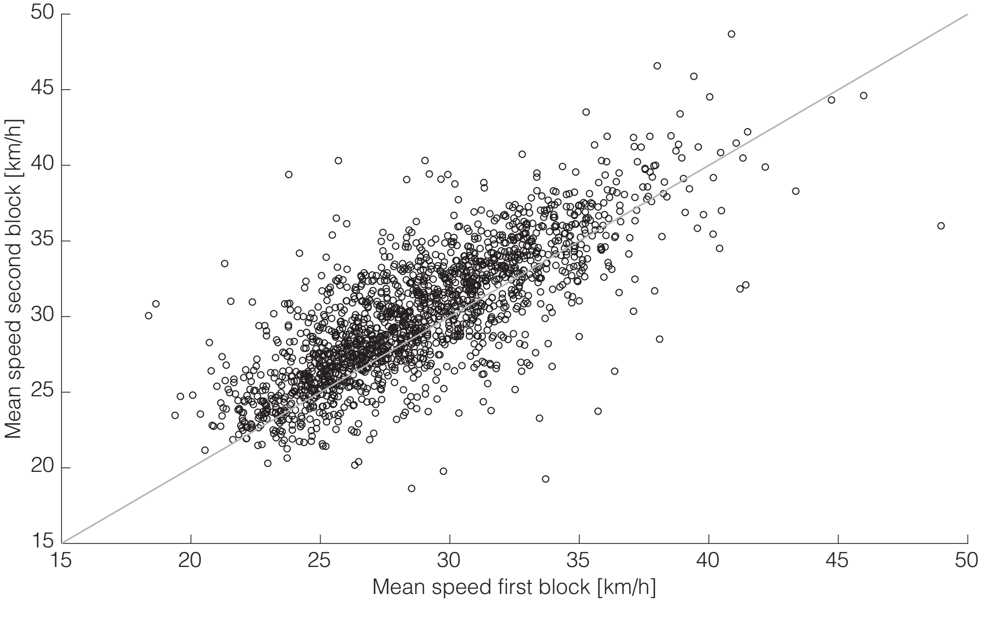


Figure 1. Students' (*N* = 1,760) mean speed of two subsequent laps around a square block of intersections. The line of unity is depicted for reference.

sponsible for their own task demands. An important finding was that there was a significant and strong correlation ($r = 0.76$, $p < 0.001$) between the time to complete the first block and the time to complete the second block, which suggests that speed is a robust individual characteristic.

2.2. Self-training

An experiment was conducted to compare three forms of feedback (no feedback, verbal feedback, and tactile feedback) while training lane keeping on a curved road. The experiment was conducted in a DDS (Figure 2). The lane width was 5 m.

Thirty male participants without any driving experience were randomly assigned to one of three conditions. One group ($n = 10$) drove without feedback on their lane keeping performance: The students had to use task-intrinsic feedback only. A second group ($n = 10$) was provided with verbal feedback based on their lateral deviation from the lane centre (too much to the left/right when approaching the lane boundaries). A third group ($n = 10$) received vibratory, tactile feedback from the seat bottom (Figure 2) as a function of the absolute deviation from the lane centre.

A classical pretest-posttest design was employed. The experiment consisted of three subsequent sessions: a 130 s pretest (no feedback for any of the three groups; final 15 s were removed from the analysis), a 600 s training session on lane keeping (different feedback for the three groups), and a 115 s posttest (no feedback for any of the groups, same route as in the pretest). Speed control was automated in the entire experiment; participants only had to steer. Participants were instructed prior to the training sessions by means of written handouts to drive as properly as possible within the right lane. The task instructions also stated that the experiment com-

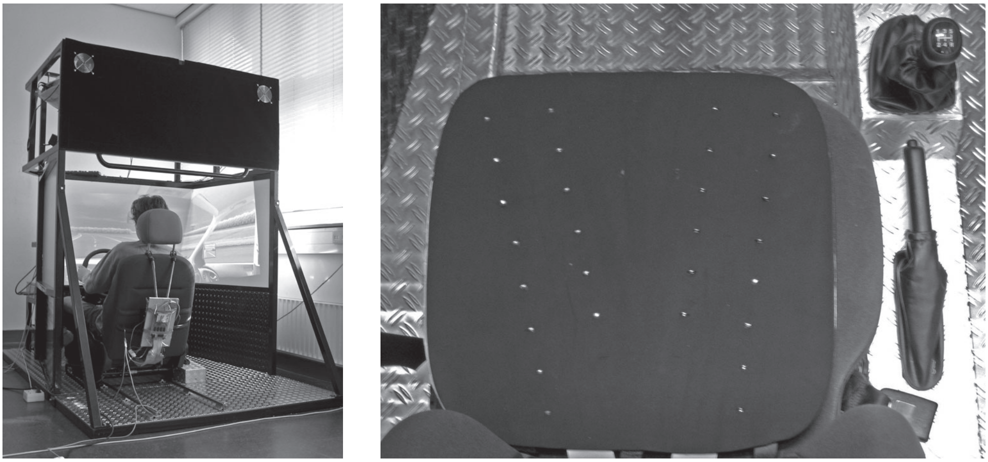


Figure 2. Dutch Driving Simulator (DDS, left) and seat bottom with vibrating elements (right).

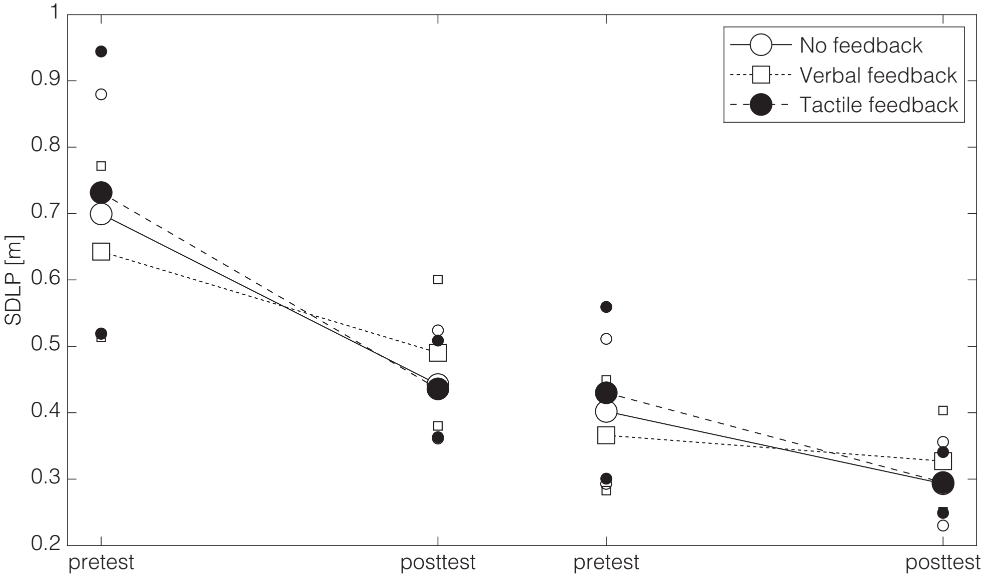


Figure 3. Mean SDLP in the pretest and posttest. The smaller markers are depicted at the mean ±1 SD (*n* = 9 for no feedback, *n* = 10 for verbal feedback, *n* = 10 for tactile feedback). Left: SDLP of the whole session, right: grand mean based on road segments' SDLP.

prised three sessions, the final of which was a driving test. The dependent measure was the standard deviation of lateral position (SDLP), a measure that has been shown to be a sensitive descriptor of lane keeping accuracy (De Winter et al., 2006d). Here, a distinction was made between the SDLP of the whole session and the average of the SDLPs of each road segment (i.e., each straight and corner separately).

The data of one participant from the *no feedback* group were recorded incorrectly and therefore not used in the analysis. The results (Figure 3) showed that, although there were large individual differences in pretest and posttest performance, learning had occurred for all three groups. That is, participants of all groups performed significantly better in the posttest than in the pretest ($p < 0.05$ for all 6 comparisons using a paired t test). When expressing the amount of learning as the pretest-posttest SDLP difference divided by the pretest SDLP, the mean learning was 35% for no feedback, 21% for verbal feedback, and 36% for tactile feedback. These numbers were not significantly different ($F = 2.12$, $p = 0.14$, using a one-way ANOVA). For the grand mean SDLP based on road segments, learning was 25% for no feedback, 9% for verbal feedback, and 27% for tactile feedback. These numbers were significantly different ($F = 3.77$, $p = 0.037$). A subsequent Tukey-Kramer multiple comparison showed that all 95% confidence intervals were overlapping, however.

This experiment showed that participants were able to learn the (predominantly visual) lane keeping task in the driving simulator without explicit verbal or tactile feedback on performance. Tactile feedback was effective, whereas verbal feedback had a (nonsignificant) tendency to be less successful than the other two methods. A possible cause of the relative ineffectiveness of verbal feedback could be information overload. Another of our studies showed that multimodal feedback for route instructions led to improved driver performance (De Groot et al., 2006). Theories of skill acquisition predict that feedback and instruction can be useful for enhancing declarative knowledge and directing the student's attention to those aspects of the task that are important (Groeger & Banks, 2007). However, instruction and feedback can also be disruptive, because working memory limitations are exceeded, student's attention can be misdirected, or the student can become reliant on instructions (Groeger & Banks, 2007). In other words, it is important to determine in which situations students actually need feedback and instructions and in which situations they are fine by themselves. Alternatively, the feedback can be provided during less demanding circumstances, for instance after a task or training session has been completed. Once the goal of the lesson and the necessary performance criteria are clear to the student, less feedback may result in more learning.

3. The virtual driving instructor

3.1. Complexity or simplicity?

A seemingly evident way to improve training effectiveness is to enhance the VDI's intelligence so that it can better understand the student's needs for feedback and instructions. It is difficult for a computer to establish the underlying cause of a particular task failure (Kappé & Van Emmerik, 2005). Consider intersection behaviour: When a student does not give right of way to another road user, this could be because the student forgot the traffic rule, had made a mistake in the type of intersection, or had incorrect viewing behaviour (Kappé & Van Emmerik, 2005). Sophisticated techniques can be used, such as expert systems or cognitive models that aim to mimic human intelligent behaviour.

It has been argued before, however, that the key to improve training effectiveness does not lie in emulating a human instructor in real time, or having a human assistant in the simulator. Instead, the advantages of simulators should be better exploited (chapter 1; De Groot et al., 2007). Objective performance ratings of students can be used to provide consistent feedback on performance, something that is not possible in a real car with a human instructor, but which is important for effective skills training (De Groot et al., 2007).

Several caveats are in order concerning the construction of complex software. Managing both hardware and software of a training simulator is an expensive and time-consuming process (Verstegen, 2003). Therefore, a more complex computer code may be detrimental to the cost-effectiveness of a simulator. Moreover, research in the related area of on-road advanced driver assistance systems (ADAS) has shown that it is essential that the driver has a clear understanding of the system (Stanton et al., 1997); a flexible and dynamic VDI can be countereffective, because the student may fail to grasp what the normative driving criteria are. Finally, as a general principle, simple models of human behaviour are preferred over more complex models, because simple models are more easily falsified (e.g., Jacobs & Grainger, 1994). It is therefore better to limit the complexity and keep the software simple when possible.

3.2. Regressive instruction

As shown in section 2.2, it is possible to learn without feedback when the environment features sufficient task-relevant information. However, for learning more complex and less visual-based tasks such as driving away, feedback and especially instructions are considered crucial. Yet, the amount of feedback and instruction should decrease with increasing practice because the student should eventually be able to carry out a task autonomously.

Current DDSs feature automatic regressive instruction, which is a relatively simple form of student-adaptive training. When the student performs a task for the first time, he or she receives step-by-step instructions (level 1). After successfully completing the task a number of times, the student is promoted to level 2, which features only corrective feedback on students' mistakes. The third and highest level assumes that the student can act autonomously. In case that the performance of the student drops, the VDI reverts to a lower learning level (Weevers et al., 2003b).

To evaluate the effectiveness of regressive instruction in learning the driving-away task, data were collected of all students who completed a training session in a DDS in the Netherlands in the period August 2005 – March 2006. The session started with video instructions and demonstrations, after which students had to repeatedly drive away and bring the car to a full stop for 10 minutes on a straight road.

Figure 4 shows the students' learning levels as a function of attempt. 84% of the students were able to autonomously drive away within 14 attempts, indicating that the regressive instruction was successful for this task (according to the VDI's criteria for success). Moreover, students' efficiency improved with practice (not shown in graph): The task completion times decreased with attempt number (the first attempt to get the car moving lasted 31.9 s on average; the fifth attempt 18.2 s, and the tenth attempt 14.6 s). Nevertheless, 16% of the students were not in level 3 at the 14th attempt. These were regularly cases in which a student repeatedly failed the task for the same reason. Independent analyses showed that certain procedural errors such as incorrect gear selection, were not recognized by the VDI. Hence, the VDI does

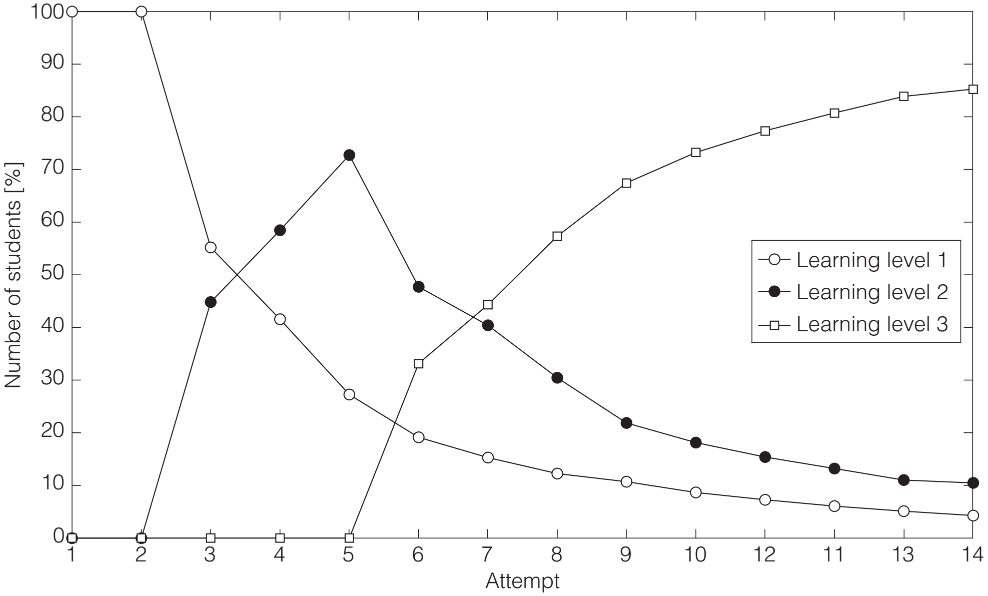


Figure 4. Number of students (percentage of sample) who performed the driving-away task at learning level 1, 2, or 3 as a function of attempt. The number of students with at least 1 attempt was 2,048; the number of students with 14 or more attempts was 1,520. The session lasted 10 minutes.

not recognize and remedy all types of failures that students make during this task, signifying the need for improvement.

It is recommended to conduct extensive user tests and software checks to investigate whether the strictness of the assessment criteria are appropriate, whether the instructions are unambiguous, whether the timing of instructions is correct, whether there are no software bugs, etcetera. In previous research, it has been suggested that such changes may lead to considerable gains in training effectiveness (chapter 9). This is not student-adaptation as such, but rather a method to enhance overall didactic quality.

It is concluded that regressive instruction based on past performance – a relatively simple form of VDI adaptation – can be effective in letting students drive away autonomously. There are indications that gains in training effectiveness can be achieved by optimizing the quality of feedback and instructions (thereby addressing all students). Improving the intelligence of the VDI should be done only with careful consideration, as the associated increase in complexity may ultimately harm training effectiveness.

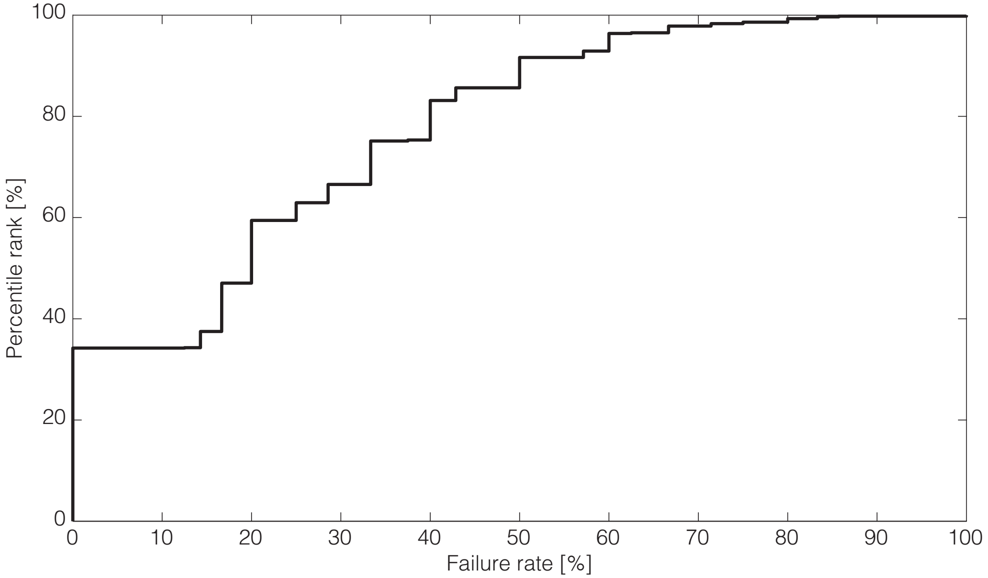


Figure 5. Percentile rank versus failure rate concerning 'pressing the clutch too early when stopping in front of a stop sign' based on a large group of students. For example, when a student fails in 2 out of 10 tasks, the failure rate is 20%, which corresponds to a percentile of 47%, a mediocre performance (adapted from De Winter et al., 2006b).

4. The student-profile

This section investigates how the study of individual differences can be used in constructing a student-profile. The idea is that a student-profile can be an important guide in decision-making about which lessons the student should follow, when to transfer to a real car, or in predicting whether someone is an accident-prone driver or not. The student-profile can be constructed from driver performance and/or from individual characteristics (e.g., age, gender, personality).

4.1. Norm-referenced assessment

A previous analysis of driver simulator training performance records has shown that task difficulty is uneven between tasks and between sessions (De Winter et al., 2006b). This means that a student's task score does depend not only on competence, but on average, it is a function of the software and strictness of the assessment criteria. Norm-referenced assessment was proposed as a solution. To be precise, the student's performance is transformed to percentiles relative to all students who had previously completed the same training programme (De Winter et al., 2006b). Figure 5 illustrates this principle based on performance records of a particular task (stopping in front of a stop sign) in a particular session. Norm-referenced assessment is a regular procedure in training and testing. The IQ test is likely the best known example (having a mean of 100 and a SD of 15 points).

As a spin-off of the normalization principle, present DDSs provide norm-referenced grades on a scale from 1 to 10 on a matrix of task-session combinations (Green Dino, 2007). These so-called strength-weakness reports are used to assist human supervisors in deciding which driving lessons should be completed by the student. An auxiliary spin-off has been the calculation of student's mean grade on all tasks. This opportunity is currently being exploited in an online driver training competition (De Groot et al., 2007; Green Dino, 2008). At present, more than 2,500 students have participated on a voluntary basis in this online ranking to compete for prizes such as free driving lessons. One supposed key advantage is that this competitive component improves the motivation to perform as well as possible, within all the traffic rules. A recent study from Sweden found a positive correlation between experience with computer games and skill-oriented aspects of car driving, whereas there were no negative effects on attitude-oriented variables (Backlund et al., in press).

A drawback of norm-referenced assessment is that the progress of a population cannot be measured. We stress that the population performance should not be used as a substitute for the norm. In the end, every driver should be able to safely carry out the crucial driving tasks. Nonetheless, norm-referenced assessment allows getting an objective indication of where the student stands with respect to the population. This is impossible during training on the road because performance is not stored into a database.

4.2. Driver assessment using factor analysis

As a follow-up study of the normalization process, we investigated whether the statistical method factor analysis can be used for driver assessment (chapter 3). Factor analysis is a technique that goes beyond the scores of individual tasks in order to reveal the underlying latent structure. To illustrate, Figure 6 shows the number of times a student had the failure *driving too fast* versus the number of times the student had the failure *following too closely*. To prevent any causal physical relationship, *driving too fast* was counted for even training sessions and *following too closely* for odd sessions. A positive correlation ($r = 0.58$, $p < 0.001$) exists between these two variables, indicating that they are partly redundant and possibly governed by the same common factor (i.e., a student's tendency for violations).

Factor analysis uses the matrix of correlations amongst a great many variables to extract a small number of factors that explain the driver behaviour. Using this methodology, a speed-score, error-score (or inversely: accuracy), and violation-score have been calculated for each student based on their task scores and mean task completion times (chapter 3). The factor scores showed predictive validity with re-

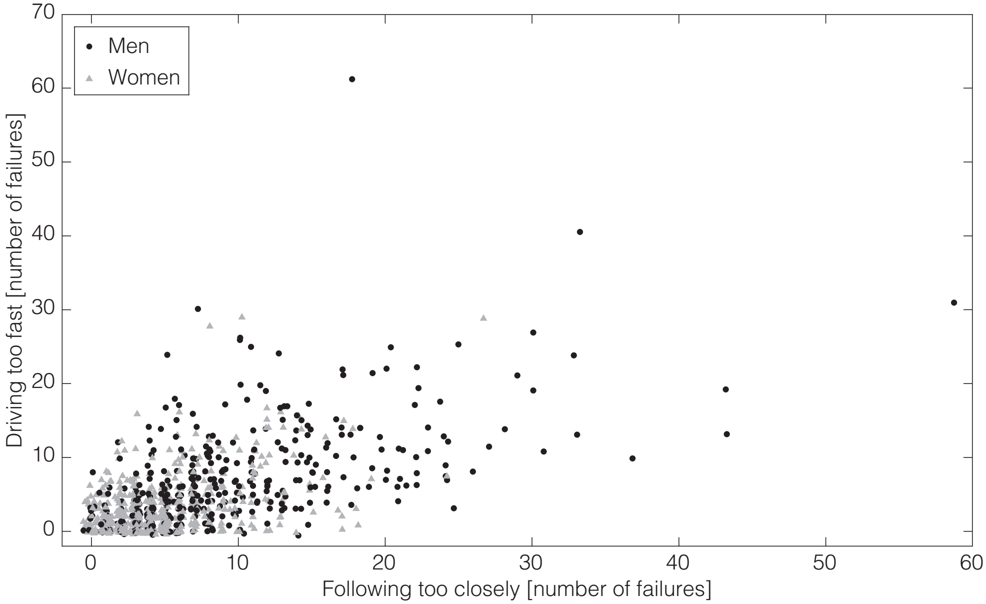


Figure 6. Number of times a student had the failure *driving too fast* versus the number of times the student had the failure *following too closely* for 517 women and 393 men who completed all sessions of the 7.5-hour training programme. A random noise with standard deviation of 0.20 has been added to prevent that dots lie exactly on top of each other.

spect to the results of the on-road driving test (chapter 4). Correcting those students with a high violation-score is particularly important for road safety as research has shown that deliberate violations (rather than errors) are predictive of road crashes (Parker, 2007).

4.3. Individual characteristics

This section explores the effects of age, gender, and personality characteristics on performance in driving simulators.

It is well established that young drivers have greater tendencies towards risk factors such as speeding, going out at night, sometimes in combination with alcohol or drugs (OECD, 2006). Older drivers, on the other hand, tend to have lower physical and mental capabilities than younger drivers (Stelmach & Nahom, 1992). A number of studies have shown that, in the simulator, young experienced drivers adopt higher speeds and make fewer errors than older drivers. Correlations typically varied in the range of -0.3 and -0.6 between driving speed/performance and age, indicating that a simulator shows sensitive age effects (Allen et al., 2007b; De Winter et al., 2006a; H.C. Lee, 2003). However, correlations were considerably smaller in magnitude (up to -0.2) during pre-licence simulator training (chapter 4), most likely due to a restriction of range: virtually all students engaged in simulation-based training for obtaining their driving licence were between 17.5 and 30 years.

On the roads, men drive more than women and are more involved in risk factors such as speeding and sensation seeking (OECD, 2006). In the Netherlands in 2007, 64 men between 18 and 24 years of age died in car crashes as compared to 10 women (other modes of transportation than cars excluded) (SWOV, 2008). Women, on the other hand, are more likely to be involved in errors and accidents that are related to operational control such as low-speed manoeuvring (Maycock & Forsyth, 1997). Gender differences during simulation-based driver training were generally large, up to 1.2 standard deviations (chapter 3; chapter 4). On average, men made fewer (steering) errors, had lower mean task completion times (a higher speed-score), and made more violations (e.g., speeding, following too closely) than women (see also Figure 6). There were no indications that the simulator was unfair. Learning rates were comparable as well as correlations with driving test results (chapter 3; follow-up study of chapter 4).

As part of previous work, we established significant correlations between several personality scales (e.g., Driver Behaviour Questionnaire (Mesken et al., 2002), Sensation Seeking Scale (Horvath & Zuckerman, 1993), and Big Five personality traits) and performance in the simulator (speed, safety-margins, lane keeping accuracy). Correlations between these individual predictors and driver performance were always lower than 0.50. This means that a certain share (< 25%) of the variance is explained, but the majority of the variability of behaviour cannot be explained by a single predictor. The most powerful predictor that we have found in the literature was an intelligence test, which predicted future flight training duration with a correlation of -0.6 (Spiker et al., 2007). An overview of cognitive capacity diversity in relation to computer task performance is provided in Neerincx et al. (2005).

Hence, it seems feasible that a student-profile, incorporating a *combination* of individual characteristics can explain a substantial share of the variance of (future) driver behaviour in the simulator or on the roads. The student-profile can potentially be used for summative and formative assessment and for student-adaptive guidance. More research is needed to investigate the potential benefits of a factor-score-, gender-, age-, personality-, and/or intelligence-differentiated training programme.

5. Impact on research and industry

This chapter provided recommendations for advancing cost-effective, student-adaptive initial driver training.

Results show that (a) The key to a successful virtual driving instructor (VDI) is to determine in which situations the students actually need feedback and instructions. (b) It is recommended to be careful with respect to increasing the VDI's complexity as this could be countereffective. Regressive instruction based on past performance – a relatively simple form of VDI adaptation – was successful in letting most students drive away autonomously. Future developments should be directed towards optimization of feedback and instructions. (c) A student-profile that incorporates a combination of individual characteristics can likely explain a substantial share of the variance in driver behaviour. The remediation of deviant behaviour of those with a high violation-score is particularly relevant to road safety. Further research is needed to evaluate how the training should be tailored towards these individual differences. A strength-weakness report based on the research shown in section 4.1 is currently used in Dutch Driving Simulators.

A remark is made with respect to data storage. The present study drew heavily on the analysis of large amounts of data stored in simulators across the Netherlands. Remarkably, in a survey of simulators (on military sites) it was found that only very few simulators had facilities for long-term data storage (Verstegen, 2003). This makes it impossible to evaluate training effectiveness of different forms of feedback and instruction. We therefore stress the importance of data storage facilities and research for evaluating and improving the effectiveness of a driver training system.

CHAPTER 6

Driving simulator fidelity and training effectiveness: A literature study of stereo presentations

Abstract

This chapter elaborates on the relationship between driving simulator fidelity and training effectiveness. The use of visual stereo presentations in driving simulators is used as a case study. A literature review indicates that a stereo presentation is associated with advantages and disadvantages. Advantages are that it can provide a relevant cue for near-distant driving tasks, induce positive reactions amongst participants, improve validity and credibility of data, improve performance and learning, and create new possibilities for augmented stereo instructions. Disadvantages are that it can increase costs, simulator sickness and distraction, and induce performance reduction in case of display artefacts. Furthermore, various complicating factors related to human perception and individual differences make it difficult to predict the effects of fidelity on training outcome. It is concluded that fidelity requirements are dependent on a compromise; the mentioned advantages and disadvantages of an intervention should be carefully weighted.

De Winter, J.C.F., Wieringa, P.A., Dankelman, J., Mulder, M., & Van Paassen, M.M. (2007b). Driving simulator fidelity and training effectiveness. *Proceedings of the 26th European Annual Conference on Human Decision Making and Manual Control*, Lyngby, Denmark. (adapted with minor textual changes)

1. Introduction

Simulators are increasingly used for initial driver training. Per definition, a (driving) simulator offers merely a representation of reality. That is, a simulator reproduces the states, behaviours, and perceptions of the real world to a limited degree. The inherent limited fidelity of a simulator is frequently considered as a drawback compared to driver training on the roads. Yet, it has been reported that further developments of driving simulators should not be aimed at increasing fidelity, but at the improvement of didactics and courseware instead (e.g., Kappé & Van Emmerik, 2005; Vlakveld, 2005b). This chapter aims to give insight into the role of fidelity with respect to the effectiveness of simulation-based driver training: Should research efforts be directed towards (increasing) fidelity or not?

Many driver-training simulators are of a medium-fidelity level, not unlike the one shown in Figure 1. This particular simulator has a relatively large horizontal field of view, provides force feedback on the steering wheel but has no enhancements such as a moving base or stereo imaging. Stereo presentations are regarded a technically feasible and promising feature for driving simulation (e.g., Balogh et al., 2006; Kemeny, 2000) and will therefore be used as a case study.

In this chapter, first, general insight is obtained into the intricacy of driving simulator fidelity. Then, the case of stereo presentations is addressed. More precisely, it is evaluated to what extent stereoscopic cues are valuable for real driving and their effect on participant's reactions as well as on task performance and learning. Finally, a summary, conclusions, and recommendations are provided.

2. Fidelity

This section reviews several unknowns of fidelity and simulation-based training.

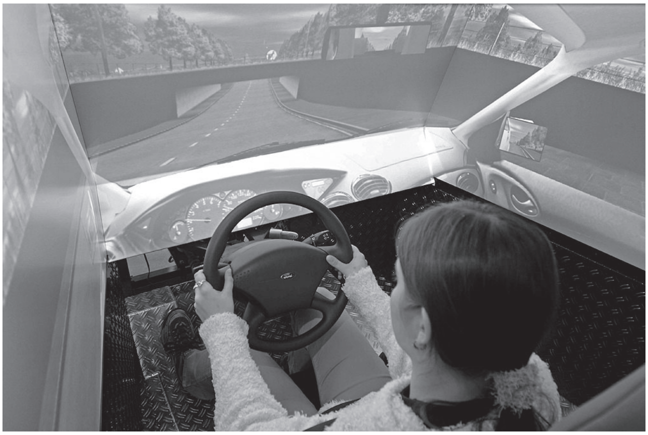


Figure 1. Driving simulator used for initial driver training (Green Dino, 2007).

2.1. Definitions of fidelity

Roza (2005) provided an overview of fidelity theories and stated that qualifying and quantifying the level of fidelity is "an area in which there exist many incomplete, inconsistent and widely scattered views, concepts and approaches" (p. ix). Different expressions are encountered in literature, such as physical fidelity, objective fidelity, perceptual fidelity, behavioural fidelity, functional fidelity, attribute fidelity, abstract fidelity, psychological fidelity, and concrete fidelity. In this study, fidelity is defined as realism or faithfulness of the simulation in the broadest sense.

2.2. Level of fidelity

The main reasons for pursuing higher levels of fidelity are based on the assumption that higher fidelity improves validity of performance, and that skills learned in a high-fidelity simulator transfer more successfully to later on-road driving. The downside is that, generally, higher fidelity increases costs (e.g., AGARD, 1980; Roza, 2005).

Different views exist on fidelity requirements. Some imply that higher fidelity is better. For example, Tidwell (1990) conducted a study on the capabilities of stereo presentations in (flight) simulation and concluded: "As high quality stereoscopic simulators are put into use, more effective training and research will result" (p. 582). Others state, more or less conservatively, that successful transfer of training does not require high-fidelity simulators per se (e.g., AGARD, 1980). Some researchers emphasize advantages of low-fidelity simulation or deliberate deviations from reality. According to J.D. Lee (2004), the pursuit of higher levels of fidelity may be inadequate because it undermines experimental control, limits data collection, dilutes training, and increases simulator sickness: "In fact, low-fidelity simulators or simulators that intentionally distort the driving experience may be more effective than those that strive for a veridical representation of the driving environment and vehicle dynamics" (p. 2253).

2.3. Consequences of limited fidelity

It is unclear whether the inherent limited fidelity of a driving simulator undermines its effectiveness and what kinds of tasks would be most affected. According to Kappé and Van Emmerik (2005), because of limited fidelity, current simulators do not allow training of (perceptual motor) vehicle control, social interaction, and complex traffic participation. Vlakveld (2005b) questioned whether driving simulators produce virtual environments that are rich and varied enough for the acquisition of higher order skills, such as situation awareness and risk perception.

In contrast, a European Commission report elaborated on driver training and simulation and stated: "In accordance with more modern thinking on the use of simulators for driver instruction, the driving simulator is regarded as a tool for promoting risk awareness and as a way of allowing the student to try out various driving situations which cannot be planned in regular traffic or which would by nature involve excessive danger on the road" (Hoeschen et al., 2001, p. 30). Welles and Holdsworth (2000) provided statistical and anecdotic support that low-cost simulators reduce accidents and improve safety awareness of police officers. Flipo (2000) elaborated on the training capabilities of a truck driving simulator and stated that it is very favourable for training manoeuvres, suggesting that it can be used for more than two thirds of the training, and that it is twice as time-efficient as a real vehicle.

2.4. Fidelity and behaviour

The effects of changes in fidelity on driving behaviour are still being explored by the scientific community. Recently, Boer (2006) analysed driving performance of participants who performed a standard car-following task in four simulators in different configurations. Results indicated that very large intersimulator differences existed in mean following distance and pedal control effort. These differences could be explained from the (chosen) car dynamics and visual contrast between the lead vehicle and the road surface. Remarkably, Reed and Green (1999) reported to have found no important differences in driving performance between a high detail colour display and a display comprising nothing more than white road-edge markings on a black background. A review by Kemeny & Panerai (2003) on human perception in driving simulation indicated that although past experiments have provided many insights, yet clearly many questions remain unanswered about how simulator characteristics affect driving behaviour.

3. Relevance of stereopsis to car driving

Driving is a task for which the visual system plays a dominant role. Perceiving one's own motion and the (relative) velocities of and distances to other vehicles and the road limits are crucial. Humans employ at least nine visual cues for perceiving depth (Cutting, 1997). Monocular cues, such as occlusion and height in the visual field, require one eye only. Binocular stereopsis results from the fact that a slightly different view of the world is projected onto each retina. The human brain integrates both projections to obtain information on depth and distance. In addition, by virtue of stereopsis, the two eyes point inward and focus on the same object yielding oculomotor cues called convergence and accommodation. Stereopsis is a complex phenomenon, covering a multitude of research topics such as its neurological aspects, interaction with monocular cues, learning-like changes of stereopsis with practice, and stereoscopic vision under dynamic conditions.

Mollenhauer (2004) reports that stereopsis can have a positive effect on performance, provided that the driving task relies on information within the distance range for which stereopsis is effective. Most distance judgments during car driving are reported to be in the range 5–500 m (Evans, 2004), whereas average intervehicle spacing during car-following typically ranges between 7 m (at 10 km/h) and 30 m (at 120 km/h) (Piao & McDonald, 2003). It is possible that drivers extract depth information from the near-distant roadway scene (Mollenhauer, 2004). Figure 2 serves to illustrate that the road can be perceived from a minimum distance of 6 m from the driver's head when looking through the front window. Visibility varies according to the seating position and the type of car (Vargas-Martín & García-Pérez, 2005).

Sachsenweger and Sachsenweger (1991) indicate that stereopsis is essential in road traffic within a range up to 20 m. Kemeny and Panerai (2003) report that stereopsis is effective for near space distance cueing (i.e., inside the vehicle or in its close vicinity), but its effectiveness more distally (e.g., for observation of other vehicles or makings on the road) is controversial. Bauer et al. (2001) concluded that stereopsis has a positive effect on driving performance only in dynamic situations at intermediate distances.

Westlake (2001) reviewed the effects of monocularity by analysing the case whether it should be safe for a monocular racing driver to participate in motor races, a task for which good vision is critical to safe operation. Westlake concluded that the one-eyed individual has deficiencies with respect to the visual field, stereopsis, and the

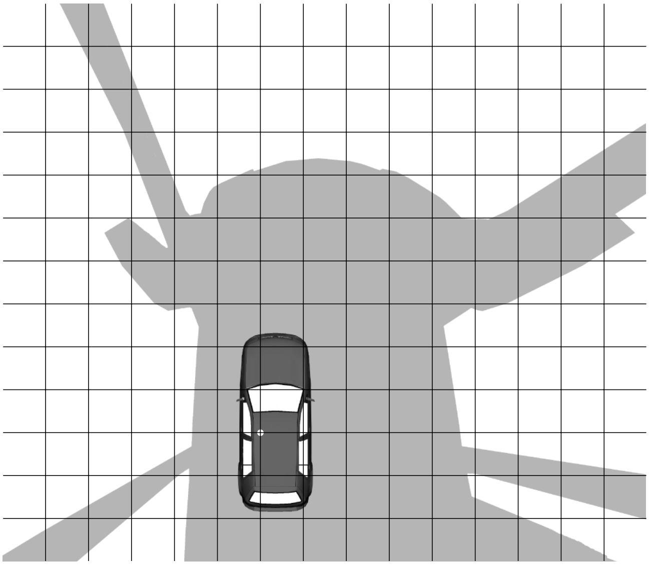


Figure 2. Area on the road surface that is not visible through the windows of a typical car (Volkswagen Golf IV) is shown in grey. Cells are 1 x 1 m. The figure was created by placing two light sources at the approximate eyes' locations in a three-dimensional model of the car. An intraocular distance of 65 mm was assumed.

ability of maintaining vision under temporary blindness in one eye. However, humans can adopt effective adaptive strategies to compensate for these deficiencies, so that the functional significance of the remaining disability is open to debate. Note that monocularity encompasses more than stereoblindness alone (such as reduced field of view and reduced contrast sensitivity). To recapitulate, literature is indecisive regarding the extent to which stereopsis is relevant to car driving.

4. Reactions to stereo presentations

Emulating stereo implies presenting disparate images to each eye, which can be accomplished with the aid of optical devices, such as head mounted displays, or with novel applications that do not require goggle devices, such as autostereoscopic or holographic systems.

A stereo presentation is often associated with positive reactions, an increased sense of realism, and increased presence, that is, sense of actually "being there" (e.g., Mollenhauer, 2004; Nash et al., 2000). However, there are unknowns about the consequences of increased presence. It has even been reported that higher presence can reduce learning due to cognitive overload (Whitelock et al., 2000).

Stereo presentations are frequently associated with an increase of simulator sickness, which could affect an operator's performance in various negative ways, such as loss of motivation, avoidance of tasks that are found disturbing, and distraction (Mollenhauer, 2004). An improved visual scene is no guarantee for elimination of simulator sickness. Instead, simplifying the visual display was hypothesized by Luke et al. (2006) to reduce simulator sickness.

Distraction might also result from the use of glasses or head tracking devices, or from irrelevant-to-training-task use of stereoscopic cues. Distraction may reduce training effectiveness (Parkes, 2005) and can reduce sensitivity of performance assessment as well. Blaauw (1982) reported that performance in a simulator was a more sensitive discriminator of driving experience than performance in an instrumented car on the road (i.e., maximum realism), probably because a simulator provides less redundant information.

Some stereo presentations induce cue conflicts or artefacts. The accommodation-convergence cue conflict is well known to occur with shutter glass systems and head mounted displays, resulting in eyestrain. Head mounted displays are reported to be excellent candidates for driving simulation as they combine movement parallax with stereo vision. However, timing discrepancies between head movement and the visual image have to be considered, especially when using a motion platform (Kemeny, 2000). Shutter-glasses might result in annoyance, poor lighting, flicker, and interocular crosstalk (Breedveld et al., 2000). Several techniques exist, aimed at reducing ocular discomfort (Eichenlaub, 2007; Mollenhauer, 2004) or to provide a conflict-free view using holographic imaging (e.g., Balogh et al., 2006). Lambooij et al. (2007) provide compromising rules of thumb for comfortably viewing stereo presentations.

Cutting (1997) warned that individual differences in the use of depth cues are large, which raises the challenge of predicting these individual differences. Literature confirms that variation in ability to distinguish between disparate images is large: Prevalence of stereo impairment amongst young observers varies between 2 and 30%, dependent on methodology (Zaroff et al., 2003). Cutting (1997) compared adjusting the image projection of a simulator to taking pictures with different lenses: when perceiving a different projection or print, the human will not overtly notice changes, yet there may be marked effects on how the scene is perceived. He particularly warned for stereopsis being a malleable cue: It is easy to provide an unrealistic illusion of depth.

5. Effects of stereo on performance and learning

Stereo presentations have frequently been reported to improve task performance in simulated environments (e.g., Kim et al., 2005; Mollenhauer, 2004). Several studies yielded mixed results (Mollenhauer et al., 2004; Nash et al., 2000) or showed that a stereo presentation can reduce performance when display artefacts are present (Pfautz, 2001).

Modest evidence could be found indicating that stereo presentations are beneficial for training. For example, stereo presentations reduced the required training time of a teleoperating task (Drascic, 1991). Mourant and Parsi (2002) found that people who trained a pick-and-place task in a stereo environment performed better on one aspect of a real world posttest than participants who had trained in a monoscopic environment. An experiment of Luursema et al. (2008) found that computer-implemented stereopsis provided a small positive effect on learning of anatomy in the medical field. The authors considered that more research was needed before advising to implement stereopsis-enabling hardware in medical settings. Merritt and CuQlock-Knopp (1991) found that the use of stereo video improved participant's reported sensitivity to monocular cues of terrain hazards during off-road driving. However, Johnson and Stewart II (1999) found no benefits of 3D helmet mounted displays on the acquisition of spatial knowledge compared to a stationary wide screen display.

Stereo presentations create new possibilities for presenting augmented cues in stereo. For example, stereo presentations are mentioned as training tools for fighter pilots for visualizing spatial relationship of air intercepts (Mowafy & Thurman, 1993).

6. Summary of results

Results are summarized below using the following taxonomy:

+ Possible advantages for training effectiveness.

- Possible disadvantages for training effectiveness.

0 Complicating factors making it more difficult to assess training effectiveness.

Higher fidelity

+ Is supposed to improve transfer of training and validity of data.

- Is associated with higher costs.

0 Is associated with many unknowns: 1) There exist many definitions and approaches to fidelity. 2) It is unclear whether higher or lower fidelity improves driver training effectiveness. 3) No agreement exists on whether current driving simulators support or undermine the training effectiveness of lower level skills (e.g., perceptual motor control) and/or higher level skills (e.g., situation awareness). 4) The effects of changes in fidelity on driving behaviour are not yet well understood.

Relevance of stereopsis to car driving

+ Is a valuable cue for in-vehicle or near-distance tasks.

0 Is a complex phenomenon and no consensus exists about the extent to which stereopsis is a relevant cue for car driving. Amongst other factors, its value depends on the individual, the type of task, seating position, type of vehicle, and the interaction with static and dynamic monocular cues.

Reactions to stereo presentations

+ Lead to positive reactions, increased sense of realism, and increased presence.

+ Improve validity and credibility of data and feedback on performance.

- Induce cue conflicts and artefacts.

- Distract the student.

- Increase simulator sickness.

0 Have effects that are difficult to predict, because individual differences are large and humans are adaptive in using depth cues.

0 Improve presence, but the effects of presence on training effectiveness are unclear.

Stereo presentations and learning

+ Generally improves task performance of (near-distance) tasks.

+ Is modestly associated with improved learning.

+ Creates possibilities for augmented feedback and instruction in stereo.

- Could degrade task performance when display artefacts are present.

0 Is associated with unknown effects. Several studies show mixed results.

7. Conclusions and recommendations

This study aimed to give insight into the role of fidelity on the effectiveness of simulation-based driver training. Results indicated that fidelity, and more specifically stereo presentations, are associated with multifactorial effects and many scientific unknowns.

Fidelity requirements are dependent on a delicate compromise and should be carefully evaluated by weighting advantages and disadvantages of the items in section 6. Improving the force-feel characteristics of pedals is an example that may improve training effectiveness: It addresses a task-relevant cue and is likely to lead to positive reactions and more realistic driving behaviour. Negative effects such as simulator sickness and distraction seem unlikely to occur.

As its effects are often unknown or contradictory, striving towards higher fidelity is not the most obvious means for improving simulation-based training. Exploiting software-related didactic advantages of driving simulators, such as free control over the training environment or the accurate and ecologically valid performance measurements (e.g., Hoeschen et al., 2001; Lew et al., 2005), should be considered as well. The complexity of a topic is no reason to conclude that it deserves no attention. It is recommended that future research in the domain of simulation-based driver training should focus on both didactic aspects and on fidelity. The framework of advantages and disadvantages in section 6 can support decision-making in the development of driving simulators. Eventually, the results have to be validated by means of an experiment that investigates transfer of training from the simulator to the road.

CHAPTER 7

The fun of engineering: A motion seat in a driving simulator

Abstract

This study evaluates the use of a motion seat in a fixed-base driving simulator. Sixty experienced drivers participated in a braking experiment and a cornering experiment in a between-subjects design. In the braking experiment, motion seat acceleration cueing versus motion off was evaluated. In the cornering experiment, we evaluated motion cueing according to the engineering way, the fun way, and motion off. When driving under the engineering way, the driver's body was tilted outward in the corners, to simulate the forces acting on the body during driving in a qualitatively correct fashion. The fun way tilted the body in the opposite direction, into the corner, as is done in many amusement rides. Results of the braking experiment showed that the motion seat resulted in smaller vehicle decelerations, more consistent stopping positions at a stop line, and smoother braking onset compared to motion off. Results of the cornering experiment did not show any significant differences in driving performance between the three conditions. Results of a questionnaire showed that participants rated fun cueing as more realistic/satisfactory than motion off. Individual differences were large compared to the effects of the motion seat. It is concluded that the motion seat was effective in inducing more realistic braking behaviour, and that the fun cueing algorithm resulted in an improved subjective experience compared to no motion.

De Winter, J.C.F., De Groot, S., Mulder, M., & Wieringa, P.A. (2007a). The fun of engineering: A motion seat in a driving simulator. *Proceedings of the Driving Simulation Conference North America*, Iowa City, IA. (adapted with minor textual changes)

1. Introduction

Depending on fidelity, drivers in a simulator have difficulty with estimating distances, speeds, and accelerations, which can result in control strategies that differ from reality (e.g., Boer et al., 2001). In this respect, the lack of motion cues in fixed-base simulators is suggested to contribute to unrealistic behaviour (Greenberg et al., 2003). However, full motion simulators are often considered financially unattractive in the domain of commercial simulation-based driver training. Moreover, even with sophisticated simulators, it is physically impossible to correctly simulate all the accelerations that affect the human body during driving (Von der Heyde & Riecke, 2001). A motion seat on a fixed-base simulator may act as a low-cost alternative for providing feedback. A previous study using a motion seat in an immersive virtual reality environment (CAVE) showed that participants preferred motion to motion off. However, subjective preference and driving behaviour hardly differed between motion parameter sets (Mollenhauer et al., 2004). The present study aims to gain insight into the effects of a motion seat and potential motion cueing algorithms in a fixed-base driving simulator.

Participants ($N = 60$) completed a braking and a cornering experiment. In the braking experiment, a comparison was made between motion seat acceleration cueing and motion off. It was hypothesized that motion results in similar effects as those reported by Siegler et al. (2001), who compared a limited-amplitude dynamic simulator platform turned on and off. Siegler et al. (2001) found that the addition of motion resulted in more accurate stopping positions at a signpost, lower maximum decelerations, and lower jerk at the onset of braking. No differences were found for speed and distance to the target at braking onset.

In the cornering experiment, three fundamentally different conditions were compared: motion cueing according to the engineering way (Eng), the fun way (Fun), and motion off (Off). When driving under the Eng condition, the driver's body was tilted outwards in bends as is regularly done in motion driving simulators. Fun tilted the body in the opposite direction, into the corner. Evaluation of the engineering versus fun ways of motion cueing was initially proposed in a working paper of Von der Heyde and Riecke (2001).

The philosophy behind the engineering way is as follows: When driving through a bend in a real car, the centripetal forces from wheel-ground contact point towards the inside of the bend. The driver observes a reactive centrifugal force that seems to move the body towards the outer edge of the car. The car generally rolls to the outside as well, as the suspension system generates a counteracting moment about the centre of gravity. The roll angle in a real car, however, is different from the roll angle that is required for substituting a centripetal force with a gravity force in a simulator. So, although quantitatively incorrect, the engineering way can be considered a relatively realistic approach to motion cueing, as the steady-state forces are qualitatively correct.

The fun way acts oppositely and simulates motion by leaning inwards in bends, an approach that is used in many amusement rides. The fun way can be considered less realistic than the engineering way as the gravity force in the simulator has the wrong direction. Still, it is hard to conceive that millions of spectators per year in amusement rides are “wrong”. The fun way could provide advantages such as higher ratings of pleasure and lower ratings of simulator sickness. It has been suggested that the incidence of simulator sickness symptoms in amusement rides is far less than that of commercial engineered driving simulators (Von der Heyde & Riecke, 2001).

The present study hypothesizes that the engineering way causes higher ratings of realism, higher incidence of simulator sickness, and lower ratings of pleasure, than the fun way and motion off (see hypotheses provided by Von der Heyde & Riecke, 2001). In addition, it is hypothesized that the engineering way results in more realistic objective driving performance than the fun way and motion off. Here, performance measures of Siegler et al. (2001) will be used. These authors compared cornering behaviour in 90° bends with a limited-amplitude motion platform turned on and turned off. Siegler et al. found that motion caused participants to take wider bends. In addition, motion resulted in lower cornering speeds than motion off.

2. Method

The experiment was conducted in a medium-fidelity fixed-base Dutch Driving Simulator (Green Dino, 2007). The simulator was operated using an automatic gearbox. Figure 1 shows the motion seat (Frex Japan Trading, 2008) in our simulator. Longitudinal and lateral accelerations of the virtual vehicle served as proportional input for the angular position of the seat. A lateral acceleration of 8.3 m/s^2 corresponded to a leftward/rightward inclination of 6.2°. A longitudinal deceleration/acceleration of 7.7 m/s^2 corresponded to a forwards/backwards inclination of 4.7°. We did not subtract seat orientation from the visual presentation. The decision to not use visual compensation was based on a study described in Mollenhauer (2004), which found that participants preferred the condition that did not subtract seat orientation from the visuals.

A between-subjects design with three groups (Off, Fun, Eng) was applied as illustrated in Table 1. Longitudinal cueing was always provided according to the engineering way (Eng), that is, backward tilt for acceleration and forward tilt for deceleration of the simulated vehicle. Trial experiments were conducted with longitudinal Fun-cueing, but we considered that it provided an awkward unnatural feeling.

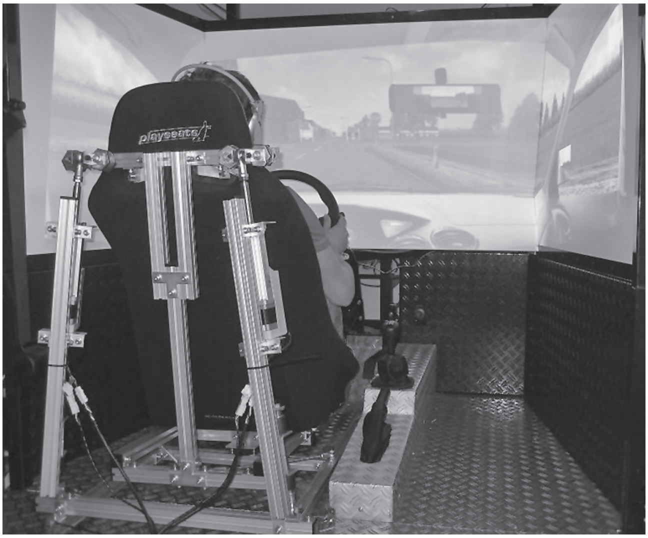


Figure 1. The motion seat that was used in the experiments.

All 60 participants were men, with at least one year of driving experience and approximately equal age (mean = 21.9 years, SD = 1.9), equal driving licence possession (mean = 3.5 years, SD = 2.1), and equal self-rated driving ability on a scale from 1 to 10 (mean = 7.7, SD = 0.91) in the three groups. The experiments were started after a 5-min learning period. All participants drove the same route and no traffic was present. The braking experiment lasted 10 min and consisted of a straight road in an urban area. At intersections the driver had to stop the car at the white line. Alternating speed limits of 30, 50, and 80 km/h were in place in-between intersections. The cornering experiment lasted 8 min and involved a two-lane rural closed track with an 80 km/h speed limit. Participants were instructed to complete both experiments with a reasonable driving style, to respect the speed limit, and to keep the vehicle within the right lane. Participants were not informed about the purpose of the experiment.

To enable a within-subjects comparison, one week after the completion of the experiment, 24 of the participants volunteered to complete a repeated measure-

Table 1. Longitudinal & lateral cueing during between-subjects experiment

	Group Off (*n* = 20)	Group Fun (*n* = 20)	Group Eng (*n* = 20)
Braking experiment (10 min)	Off	Long: Eng	Long: Eng
Cornering experiment (8 min)	Off	Lat: Fun, Long: Eng	Lat: Eng, Long: Eng

Note. Off: no motion, Fun: motion cueing according to the fun way, Eng: motion cueing according to the engineering way. Long: longitudinal cueing, Lat: lateral cueing.

Table 2. Dependent measures for the braking experiment [1]

Abbreviation	Unit	Description
NrValidStops	#	Total number of stops coming to complete standstill
Mean Vini	m/s	Mean speed at onset of braking (t = 0 s)
Mean DTLini	m	Mean distance to the target line at onset of braking (t = 0 s)
SD DTLfin	m	Standard deviation of distance to the target line when standing still (i.e., stopping consistency)
Mean max. dec.	m/s^2	Mean maximum deceleration for speed > 5 km/h (t = T s)
Mean onset jerk	m/s^3	Mean rate of change of deceleration for the first half of the stopping manoeuvre (deceleration at t = T/2 divided by T/2)

Note. The dependent measures were calculated for each participant by averaging over all stops. The first stop of each participant was removed from the analyses.*t* represents the elapsed time since the onset of braking.

ment of the braking experiment. Twelve participants drove with motion Off in the first experiment and with Eng in the second experiment. Conversely, twelve other participants drove with Eng in the first experiment and Off in the second experiment.

3. Dependent measures

Tables 2–4 show the measures that were calculated for each participant, adapted from Siegler et al. (2001). After the experiment, participants completed a questionnaire, consisting of ten questions related to realism and pleasure (see Table 4) which had to be answered on an interval scale ranging from 1 to 10 (no anchors). Participants also completed an adapted simulator sickness questionnaire (Kennedy et al., 1993). Our version did not distinguish between the nausea, oculomotor, and disorientation subscales. Moreover, the *fullness of the head* symptom was removed.[2] For each symptom, participants rated an interval scale from 1 (*no problems*) to 5 (*large*

[1]) The dependent measures have been duplicated from chapter 8.

[2]) We considered the original simulator sickness questionnaire (SSQ) not readily interpretable (e.g., how should participants interpret “difficulty concentrating” on the oculomotor scale?). Inspection of the work of Kennedy et al. (1993) showed that the three factors (nausea, oculomotor, and disorientation) were orthogonally rotated instead of obliquely, to obtain simple patterns. A general simulator sickness factor was found to explain 50% of the variance in the orthogonal solution. State-of-the-art literature on factor analysis indicates that there is little justification to using orthogonal rotation when factors correlate (Fabrigar et al., 1999). We performed oblique rotation (oblimin) of the loadings shown in Kennedy et al. (1993, p. 208). Results indicated that the three factors indeed correlate substantially (0.28 to 0.40). The rotated pattern was considered better interpretable than the orthogonal loadings. We also calculated factor score coefficients (Bartlett procedure). It was found that some coefficients were very low, which may warrant consideration omitting items from the SSQ. Based on these findings, the present authors considered it theoretically and practically justified to use an adapted (i.e., simplified and more easily interpretable) SSQ.

Table 3. Dependent measures for the cornering experiment

Abbreviation	Unit	Description
NrDepartures	#	Number of road departures
Mean LCE45deg	m	Mean lane centre error when halfway through the bend
Mean V45deg	m/s	Mean speed when halfway through the bend

Note. The dependent measures were calculated for each participant by averaging over 90° bends with radii of 15–20 m. A distinction was made between 6 left bends and 8 right bends. For each participant, the first 7.55 km were selected, equalling one lap on the closed track. When a road departure had occurred, the car was automatically placed back on the road with zero speed. Data from 10 s prior to the departure to 20 s after each road departure were removed from further analyses.

problems). A total simulator sickness score was calculated by averaging over the symptoms.

The (independent/paired) *t* test was used to statistically evaluate the difference between the means of two samples. Cohen's *d* was used as a measure for the effect size. Differences between three means were statistically evaluated using a one-way analysis of variance (ANOVA), or a Kruskal-Wallis nonparametric one-way analysis of variance.

4. Results

Results of the braking experiments are shown in Table 5. Participants completed on average about 13 stops. As expected, maximum deceleration (Mean max. dec.) was lower, stopping consistency (SD DTLfin) was better and braking onset (Mean

Table 4. Questionnaire on realism/pleasure (translated from Dutch)

Abbreviation	Question
RealDriving	How realistic did you find driving in the simulator?
RealBrake	How realistic did you find the braking?
RealAccelerate	How realistic did you find the accelerating?
JudgeSpeed	Could you well judge the speed of the car?
JudgeDistance	Could you well judge the distance to the signs in the braking experiment?
RealBends	How realistic did you find driving through bends?
JudgeSpeedB	How well could you judge the cornering speed?
RealSeat	Did the moving seat add to the feeling of realism?
Benefit	Do you think your driving performance benefited from the moving seat?
Pleasure	Did you enjoy the fact that the seat was moving?

Table 5. Results of the braking experiments

	Between-subjects experiment				Within-subjects experiment			
	Mean				Mean			
Measure	Off (*n* = 20)	Fun&Eng (*n* = 40)	*p*	*d*	Off (*n* = 24)	Fun&Eng (*n* = 24)	*p*	*d*
NrValidStops	12.4	12.4	1.0	-0.01	13.0	12.8	0.8	0.06
Mean Vini	16.3	16.0	0.7	0.12	16.0	15.9	0.5	0.16
Mean DTLini	49.2	54.9	0.2	-0.36	48.3	54.9	0.1	-0.41
SD DTLfin	2.95	2.00	0.011	0.69	2.24	1.81	0.034	0.54
Mean max. dec.	6.29	5.87	0.3	0.25	6.49	5.59	0.006	0.63
Mean onset jerk	5.99	3.14	<0.001	0.88	5.23	3.24	0.002	0.89

Note. The *p* value of the between-subjects experiment was calculated using an independent *t* test; the *p* value of the within-subjects experiment was calculated using a paired *t* test.

onset jerk) was lower for motion cueing (Eng & Fun groups aggregated) compared to Off. No significant differences were found for speeds and distances to the stopping line at onset of braking (Mean Vini, Mean DTLini). To confirm the results, the dependent measures were calculated for the cornering experiment as well, the results of which are shown in Table 6. It can be seen that motion again resulted in lower decelerations and lower jerk. In addition, participants also pressed the brake at a significantly greater distance to the bend.

Results of the cornering experiment are presented in Table 7. No significant differences were found between any of the conditions. An interesting finding was that the standard deviation (SD) of the lane centre error halfway through the left bend was considerably lower for Eng compared to Fun and Off. Closer examination was done by plotting the mean trajectories of each participant (see Figures 2 and 3). It appears from Figure 2 that the increased SDs of Fun and Off were caused by two participants who had not respected the instruction to remain on their right lane by

Table 6. Braking behaviour in the cornering experiment

	Mean			
Measure	Off (*n* = 20)	Fun&Eng (*n* = 40)	*p*	*d*
Mean Vini	20.9	20.2	0.1	0.43
Mean DTLini	45.6	49.8	0.06	-0.53
Mean max. dec.	6.58	5.26	0.003	0.82
Mean onset jerk	8.42	5.69	<0.001	1.09

Note. The measures were calculated for each participant over 15 straights on which participants braked before the bends. The measures were z-transformed before calculating the *p*-values and Cohen's *d* effect sizes.

Table 7. Results of the cornering experiment

	Mean (SD)			
Measure	Off (n = 20)	Fun (n = 20)	Eng (n = 20)	p
NrDepartures	0.70 (1.08)	0.35 (0.49)	0.50 (0.76)	0.4
Mean LCE45deg (left bends)	1.72 (1.04)	1.72 (0.78)	1.55 (0.39)	0.7
Mean LCE45deg (right bends)	-0.76 (0.37)	-0.74 (0.36)	-0.77 (0.46)	1.0
Mean V45deg (left bends)	12.1 (1.11)	11.6 (1.21)	11.5 (1.33)	0.3
Mean V45deg (right bends)	9.9 (1.25)	10.0 (1.48)	9.7 (1.50)	0.8

Note. The p value was calculated using a one-way analysis of variance (ANOVA) on the three conditions.

consistently shortcutting on the left lane. As only two participants were involved, these results cannot be attributed to the motion cueing conditions. Figures 4 and 5 show the mean speeds through the left and right bends. From Figures 2–5 it can be seen that any differences in mean trajectories/speeds were negligibly small compared to the magnitude of individual differences.

Results of the questionnaire are shown in Table 8. It was hypothesized that Fun results in a higher pleasure rating than Eng. In this study we found a pleasure rating of 6.4 for Fun and a pleasure rating of 5.9 for Eng. These numbers were not significantly different ($p = 0.4$). To prevent committing Type I errors, a multivariate ap-

Table 8. Questionnaire results

	Mean (SD)			Factor loading	h^2
Measure	Off (n = 20)	Fun (n = 20)	Eng (n = 20)		
RealDriving	4.2 (1.9)	5.3 (1.6)	5.1 (1.7)	0.88	0.78
RealBrake	3.7 (1.7)	5.7 (2.0)	4.9 (1.6)	0.75	0.57
RealAccelerate	4.7 (2.1)	5.5 (2.2)	5.7 (1.6)	0.62	0.39
JudgeSpeed	3.6 (2.1)	4.4 (1.7)	3.5 (1.3)	0.73	0.54
JudgeDistance	3.8 (2.0)	4.6 (2.2)	4.6 (1.9)	0.55	0.30
RealBends	4.5 (1.5)	4.9 (2.0)	5.4 (1.7)	0.65	0.42
JudgeSpeedB	4.1 (1.8)	4.5 (1.5)	4.6 (1.5)	0.55	0.30
RealSeat	–	6.8 (2.0)	6.5 (1.8)	0.32	0.10
Benefit	–	6.3 (1.8)	6.2 (1.8)	0.44	0.20
Pleasure	–	6.4 (2.5)	5.9 (2.1)	0.58	0.34
Factor-score	-0.40 (1.01)	0.29 (1.03)	0.10 (0.87)		

Note. The eigenvalue of the first factor was 4.46. The eigenvalue of the second, unretained factor was 1.38. A Kruskal-Wallis test indicated that the factor-scores are different ($p = 0.027$). A Tukey-Kramer multiple comparison indicated that Off and Fun are significantly different.

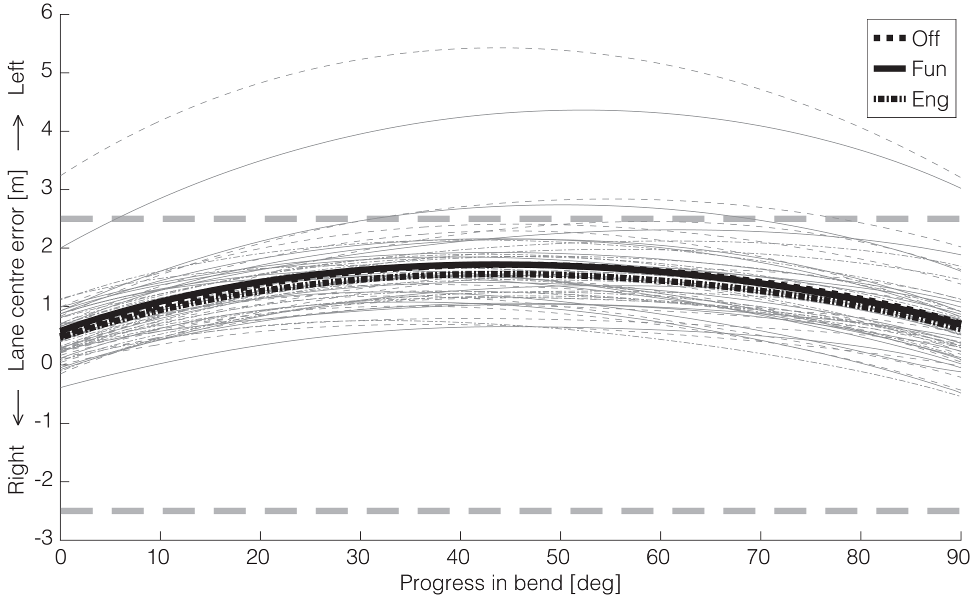


Figure 2. Mean lane centre error (LCE) in left bends. Grey lines represent participants' mean LCE of 6 bends. Lines at LCE = 2.5 and LCE = -2.5 represent lane boundaries.

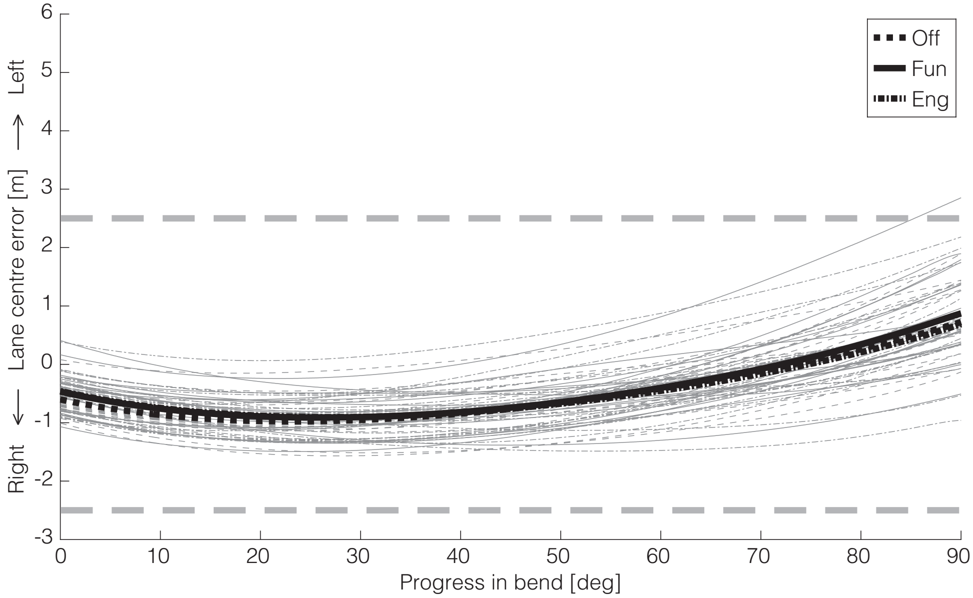


Figure 3. Mean LCE in right bends. Grey lines represent participants' mean LCE of 8 bends. Lines at LCE = 2.5 and LCE = -2.5 represent lane boundaries.

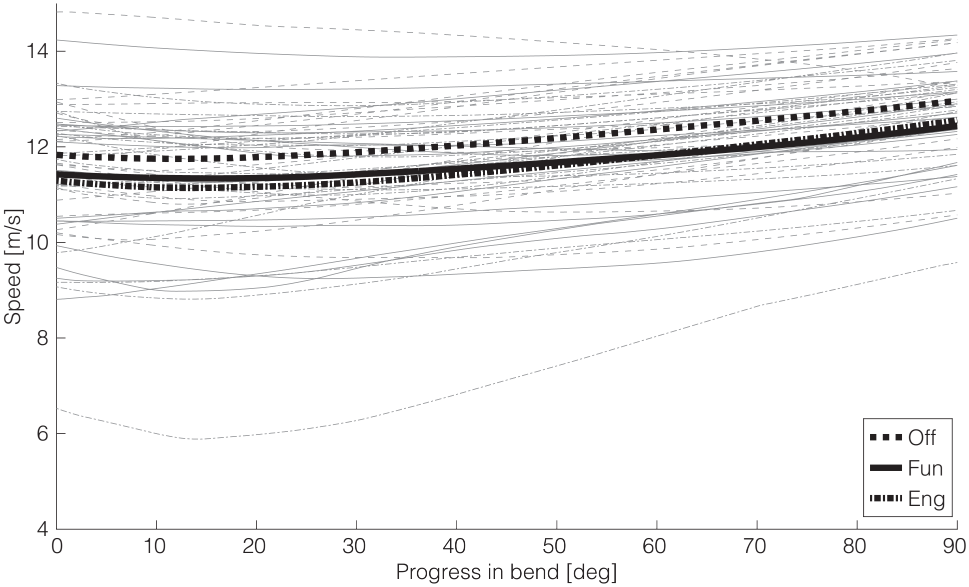


Figure 4. Mean speed in left bends. Grey lines represent participants' mean speed of 6 bends.

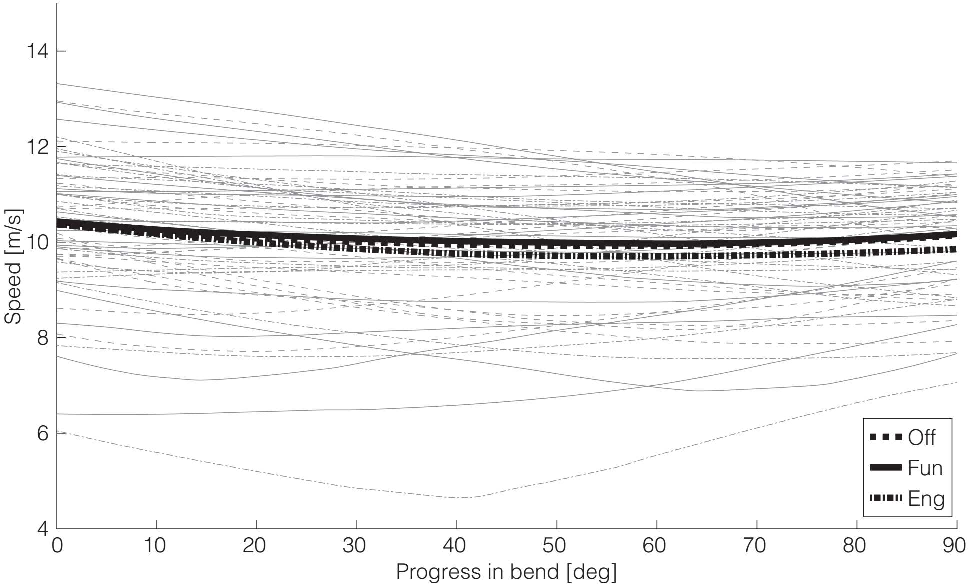


Figure 5. Mean speed in right bends. Grey lines represent participants' mean speed of 8 bends.

Table 9. Results of the questionnaire on simulator sickness

	Mean (SD)		
	Off (n = 20)	Fun (n = 20)	Eng (n = 20)
Mean sickness score	1.66 (0.64)	1.87 (0.48)	1.75 (0.60)

Note. A Kruskal-Wallis test indicated that the scores are not significantly different (p = 0.4).

proach was used on the questionnaire data. A Pearson correlation matrix was submitted to a principal axis factoring to extract one common factor explaining 39% of the variance. The decision to extract one factor was supported by the scree plot and the interpretability of the loadings compared to extracting two or more factors. Bartlett factor scores, representing total satisfactory/realism level, are shown in Table 8. A Kruskal-Wallis test indicated that the median factor scores for the three conditions were significantly different (p = 0.027). According to a Tukey-Kramer multiple comparison the satisfaction/realism score was significantly higher for Fun compared to Off.

The results of the questionnaire on simulator sickness are shown in Table 9. There were no significant differences in sickness scores between the conditions. Fun and Eng caused higher ratings on the following symptoms: general discomfort, fatigue, and eyestrain, compared to Off. After applying Bonferroni corrections these differences were not large enough to reach a critical significance level anymore.

5. Discussion

The motion seat resulted in lower decelerations, more consistent stopping distances, and smoother braking onset than Off. As expected, the results are qualitatively congruent with the effects of a limited-amplitude motion platform (see Siegler et al., 2001). It was also found that participants pressed the brake at a greater distance towards the target for motion compared to Off, whereas Siegler et al. found no effect. Although the motion seat reduced the vehicle decelerations, these were still high in comparison to on-road driving; representative values for real car driving are in the order of 3–4 m/s^2 (Siegler et al., 2001).

For cornering, no significant differences in driving performance were observed between the fun way, engineering way, and motion off. Contrary, Siegler et al. (2001) observed that participants took a wider bend and adopted lower speeds when using a motion platform. Inspection of the results of Siegler et al., however, showed that their effects were relatively small, with a Cohen's d between 0.2 and 0.3 for the speed halfway through the bend.

Participants judged Fun as more realistic/satisfactory than Off. No significant differences were observed between Fun and Eng, that is, two opposite forms of lateral motion cueing. Overall simulator sickness scores were low and not significantly different between the three conditions.

In the present study, the seat was moving without compensation of the visual scene or the pedals/steering wheel, making it possible for the participant to be aware of the seat inclination while keeping the head upright. Seat inclinations could be noticed through tactile and proprioceptive feedback, rather than through the vestibular organs as in real car driving.

The present study showed that the motion seat improved braking performance in the simulator. Further research is recommended to evaluate whether a motion seat, possibly combined with other types of motion cueing, also results in improved learning and safer driving on the roads. The seat has shown to be a successful means of improving the subjective realism of the simulator. Increased face validity may further improve students' motivation to learn. It can be expected that more realistic braking behaviour in the simulator (i.e., less abrupt braking, lower decelerations) transfers to the road, resulting in safer braking during the initial moments in real traffic. It is also recommended to further develop motion seat control algorithms. High-frequency cues may be beneficial to enhance speed and motion perception and could be particularly suited to a (limited amplitude) motion seat. Finally, it seems worthwhile to investigate individual differences. Why is it that some participants adopted much higher speeds than others? Understanding individual differences and their correlates might provide important insight for improving the effectiveness of simulation-based training.

CHAPTER 8

The search for higher fidelity in fixed-base driving simulation: Six feedback systems evaluated

Abstract

Because motion cues are lacking in a fixed-base simulator, people tend to drive faster and brake harder in the simulator than they do on the road. A motion platform is too expensive for low-cost simulation-based driver training; therefore, alternative solutions have to be sought. In this chapter, the following six low-cost motion cueing devices are tested for their effects on braking and cornering behaviour in a fixed-base driving simulator: (a) a seatbelt tensioning system, (b) a stiff brake pedal, (c) a vibrating steering wheel, (d) screeching tyre sound, (e) a vibrating seat, and (f) a pressure seat. Results indicated that all systems, except for the pressure seat, were beneficial in inducing more realistic driving behaviour. The magnitude and the nature of the effects, however, were notably different between the systems. The best results were obtained with those devices that generated stimuli of sufficient magnitude and adequate stimulus-response compatibility.

De Winter, J.C.F., De Groot, S., Mulder, M., Wieringa, P.A., & Dankelman, J. (2008c). The search for higher fidelity in fixed-base driving simulation: Six feedback systems evaluated. *Proceedings of the Driving Simulation Conference Europe*, Monte Carlo, Monaco, 183–192. (adapted with minor textual changes)

1. Introduction

Low-cost fixed-base driving simulators have regularly shown satisfactory relative fidelity such as in intra/intersubject comparisons (e.g., Lew et al., 2005). Their absolute fidelity, however, remains open to improvement. One particular problem of low-cost driving simulators is that people tend to drive too fast as compared to reality (Green, 2005). Although car driving is primarily a visual task, nonvisual motion cues are important in drivers' control of speed and heading (Reymond et al., 2001).

A disadvantage of motion platforms, however, is that they are expensive and complex systems, and therefore unattractive for cost-effective simulation-based driver training. In order to provide low-cost motion cueing in a fixed-base simulator, alternative solutions, such as a motion seat, can be used. Previous research showed that a motion seat improved face validity and realism with respect to braking behaviour (chapter 7). Nonetheless, even a relatively simple motion seat significantly increases the complexity and the cost of a simulator.

With the aim of improving absolute driving simulator fidelity, the present chapter describes six experiments carried out with different low-cost motion cueing systems. Considering that people tend to drive too fast and brake too much in fixed-base simulators, each system was hypothesized to contribute to lower decelerations during braking, lower brake onset jerk, and lower cornering speeds (see also Siegler et al., 2001 and chapter 7).

2. Systems under evaluation

Six cueing systems were developed and evaluated: (a) a seatbelt tensioning system, (b) a stiff brake pedal, (c) a vibrating steering wheel, (d) a vibrating seat, (e) screeching tyre sound, and (f) a pressure seat. The characteristics of these systems are described in this section.

2.1. Seatbelt tensioning system (longitudinal deceleration cueing)

The feedback cue was a tension force in the seatbelt, proportional to the deceleration of the virtual car. The additional tensioning forces that were applied in the seatbelt ranged from 0 to 150 N for corresponding decelerations ranging from 0 to 5 m/s^2. For decelerations larger than 5 m/s^2, the force remained at 150 N. These force settings were based on the results of a just-noticeable-difference experiment (data not shown). The seatbelt tensioning system (Figure 1) was powered by a moment-inducing motor with a reduction gearbox. The seatbelt was fed through a sleeve in a cylinder. When a moment was exerted by the motor, the belt gripped and rolled onto the cylinder, pulling the belt tightly over the shoulder and chest of the driver. When no moment was exerted, the seatbelt could be operated in regular fashion. The

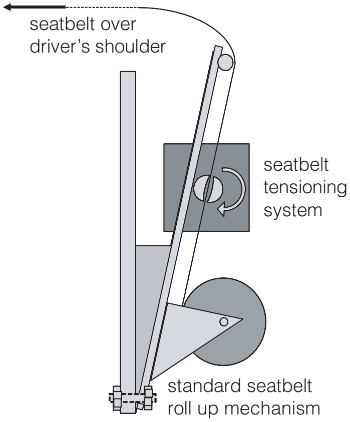


Figure 1a. Illustration of seatbelt system.

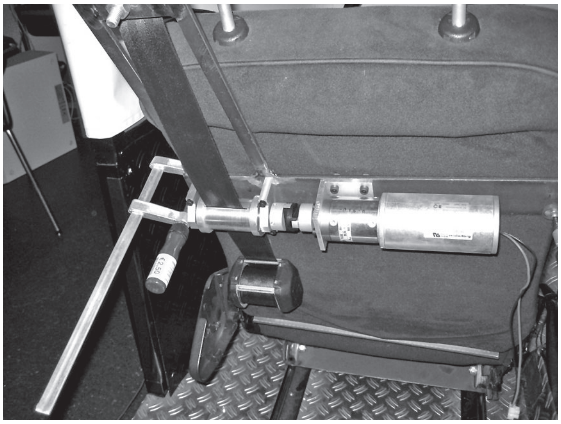


Figure 1b. Seatbelt system on the back of the seat.

design allowed the drivers to change their position during driving: No adjustments were needed for different drivers.

2.2. Stiff brake pedal (longitudinal deceleration cueing)

Additional proprioceptive cues from a stiffer pedal were hypothesized to result in improved braking behaviour because more information concerning the deceleration of the vehicle is available as compared to a softer pedal. The initial soft pedal of the simulator (stiffness about 0.06 N/mm) was modified by placing an extra spring in the brake pedal assembly to create a pedal with a common stiffness of about 1.6 N/mm.

2.3. Vibrating steering wheel (longitudinal deceleration cueing)

Vibrating elements are inexpensive, easy to implement, and safe. Transferring information through vibrations has already been successfully applied in driving simulators and in lane departure warning systems (Suzuki & Jansson, 2003). Suzuki and Jansson demonstrated that drivers respond quickly and intuitively to vibrations. In the present study, vibrations were provided by a speaker attached in the steering wheel (see Figure 2). The speaker played a low-frequency sample that became louder with increasing deceleration of the car.

2.4. Screeching tyre sound (longitudinal & lateral cueing)

Vehicle sound can increase the overall sensation of speed (Davis & Green, 1995). Relatively few simulators generate screeching tyre sound when driving near the

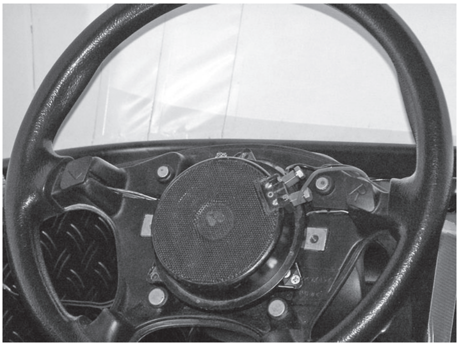


Figure 2a. Vibrating steering wheel (open).

Figure 2b. Vibrating steering wheel (closed).

performance limit of the tyres. Considering the fact that people in (fixed-base) simulators often drive unrealistically fast, screeching tyre sound could be a relevant cue for enhancing driver awareness with respect to high accelerations and high cornering speeds. The screeching sound was generated when the acceleration of the simulated car exceeded a friction ellipse, with a screeching volume proportional to the distance to the ellipse. The ellipse had a semimajor axis of 8 m/s^2 in longitudinal direction and a semiminor axis of 7.2 m/s^2 in lateral direction. Maximum grip was 9.4 m/s^2 in longitudinal direction and 8.5 m/s^2 in lateral direction.

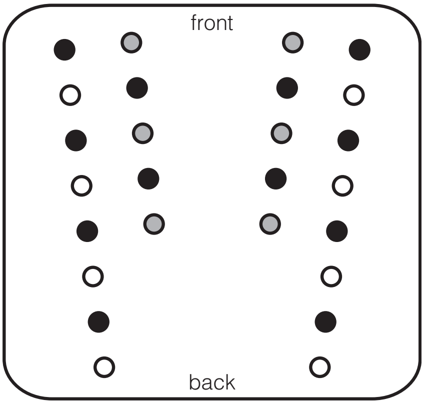


Figure 3a. Illustration of the seat's vibration elements.

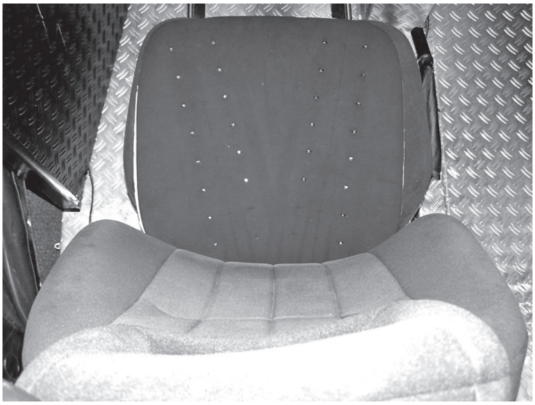


Figure 3b. Seat with vibration elements seen from above. The two centremost elements were not used.

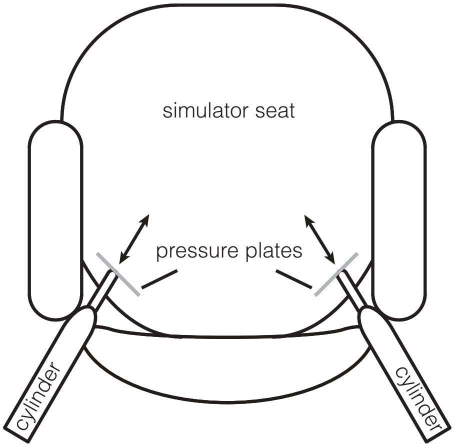


Figure 4a. Illustration of the pressure seat.

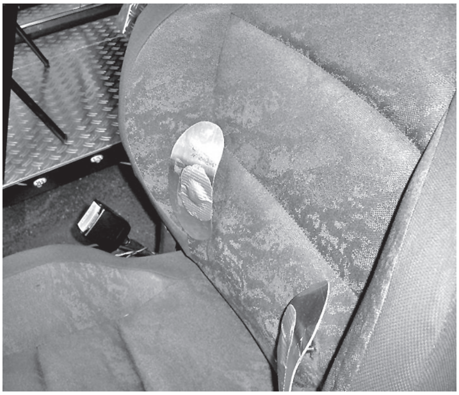


Figure 4b. Pressure plates on the seat.

2.5. Vibrating seat (lateral acceleration cueing)

To provide a sense of lateral acceleration to the driver, vibrations were provided by small DC motors with an eccentric weight (as in mobile phones) on the bottom of the seat (see Figure 3). When driving in right turns, the left part of the seat vibrated and vice versa. For lateral accelerations above 1 m/s^2, four of the outer elements (illustrated as white in Figure 3) vibrated on one side. Above 3 m/s^2, three of the inner elements (grey) also started vibrating and above 5 m/s^2 all elements would vibrate.

2.6. Pressure seat (lateral acceleration cueing)

When driving through a corner in a real car, the driver feels a reactive centrifugal force that seems to move his/her body towards the outer edge of the car. To simulate these forces in a qualitatively correct fashion, the seat of the simulator was equipped with two cylinders and an air pressure regulator (see Figure 4). The pressure of the cylinders was applied to the human body by means of metal plates, just below the ribs, this pressure being proportional to the lateral acceleration of the car. In left turns, the plate on the right side was pressurized and the plate on the left side remained unpressurized. In right turns, this situation was reversed.

3. Experiments

The experiments were conducted in a low-cost medium-fidelity fixed-base Dutch Driving Simulator (Green Dino, 2007). Table 1 provides details on the experimental protocols. All systems were evaluated in a baseline condition (Off or Soft) and an experimental condition (On or Stiff). Participants were recruited from the university community and were not informed about the experimental conditions and the purpose of the experiment.

The experiments were conducted independently from each other and therefore with somewhat different protocols. The high similarity in the way that the experiments were carried out allows a joint investigation of their performance. However, comparison of the absolute values of the dependent measures *between* experiments should be made only with careful consideration. Evidently, the conditions *within* experiments were properly matched and randomized where appropriate. Parts

Table 1. Experimental protocols

	N[a]	F[b]	E/I[c]	W/B[d]	NrRuns[e]	Duration[f]	Task[g]	Aut[h]
1. Seatbelt system	20	0	I	W	2	10 stops	A	S+G
2. Brake pedal	24	2	E+I	W	2	10 stops	A	S+G
3. Vibrating wheel	13	2	E	W	2	10 stops	A	–
4a. Screeching tyres	12	0	I	W	4	4 stops	A	S+G
4b. Screeching tyres	12	0	I	W	4	4 turns	B	G
5. Vibrating seat	15	3	E	W	4	8 turns	B	–
6. Pressure seat	31	5	I	B	1	14 turns	C	G

Note. The experiment with seatbelt tensioning system was conducted with the stiff brake pedal. The other experiments were conducted with the soft brake pedal.

[a] Number of participants in the experiment

[b] Number of women

[c] Indicates whether the participants were experienced drivers (E), inexperienced (I), or both (E+I). Inexperienced drivers were defined as drivers without a driving licence.

[d] Indicates whether a within-subjects (W) or between-subjects (B) design was applied.

[e] Number of runs per participant. For within-subjects experiments, the number of runs per experimental condition per participant was half of NrRuns.

[f] Indicates how many manoeuvres per run were taken into account in the statistical analyses. Before each experiment, participants engaged in a 4 to 10 min training run to get acquainted with the simulator.

[g] Indicates the task that participants had to perform. Three tasks were used:

A. A braking experiment on a straight road with intersections. Different speed limits of 30, 50, and 80 km/h were in place in-between intersections. Participants were instructed to respect the speed limit and come to a complete stop as smooth and as accurate as possible at the stopping lines in front of intersections.

B. A cornering experiment around a square road block. Participants had to drive either clockwise or anticlockwise, performing a 90° turn on every intersection. The speed limit was 50 km/h and the participants were asked to stay in their lane and to drive as they would normally do.

C. A cornering experiment on a two-lane rural 7.55 km closed track with an 80 km/h speed limit. Participants were instructed to drive with a reasonable driving style, respect the speed limit, and keep within the right lane.

[h] Indicates whether steering (S) or gear changing (G) was automated by the simulator.

Table 2. Dependent measures for task A

Abbreviation	Unit	Description
NrValidStops	#	Total number of stops coming to complete standstill
Mean Vini	m/s	Mean speed at onset of braking (t = 0 s)
Mean DTLini	m	Mean distance to the target line at onset of braking (t = 0 s)
SD DTLfin	m	Standard deviation of distance to the target line when standing still (i.e., stopping consistency)
Mean max. dec.	m/s^2	Mean maximum deceleration for speed > 5 km/h (t = T s)
Mean onset jerk	m/s^3	Mean rate of change of deceleration for the first half of the stopping manoeuvre (deceleration at t = T/2 divided by T/2)

Note. The dependent measures were calculated for each participant by averaging over all stops. The first stop of each run was removed from the analyses. t represents the elapsed time since the onset of braking.

of the data of the vibrating steering wheel and vibrating seat have been described earlier (Boschloo et al., 2005).

4. Dependent measures

Tables 2 and 3 show the dependent measures that were used. Earlier research showed that these measures are sensitive descriptors of braking and cornering behaviour (Siegler et al., 2001; chapter 7). Speed and distance at onset of braking (Mean Vini and Mean DTLini) are important because they effectively determine how hard the driver should brake in order to come to a stop at the stop line. The peak deceleration and the initial rate of change of deceleration (i.e., Mean max. dec. and Mean onset jerk) provide further information about *how* the driver braked. Finally, SD DTLfin provides information on the stopping consistency, that is, how precisely the driver was able to stop the car with respect to the stop line. For cornering behaviour, the nonredundant measures Mean LCE45deg and Mean V45deg provide infor-

Table 3. Dependent measures for tasks B and C

Abbreviation	Unit	Description
NrDepartures	#	Number of road departures
Mean LCE45deg	m	Mean lane centre error when halfway through the turn
Mean V45deg	m/s	Mean speed when halfway through the turn

Note. For task C, we selected only 90°-bends with radii in between 15 and 20 m; these were the most frequently occurring bends on the track. When a road departure had occurred, data from 10 s prior to 20 s after the departure were removed and not considered in further analyses.

mation on the speed and path halfway through the turns. A distinction was made between left and right turns.

The dependent measures were tested for the effects of the experimental conditions using either a paired or an independent *t* test. Cohen's *d* was used as an effect size measure. Besides the objective measures, a short questionnaire on the subjective experience and/or the subjective realism of the simulator was employed in each experiment.

5. Results

The results of the experiments are shown in Table 4.

5.1. Seatbelt tensioning system (longitudinal deceleration cueing)

The seatbelt tensioning system had a highly beneficial effect on braking behaviour. Participants' maximum decelerations were lower, they braked earlier, their stopping position consistency was better, and the braking onset jerk was lower when the system was On as compared to Off. Participants had responded to the questionnaire item whether the seatbelt improved the feeling of sitting in a real car, with a mean score of 3.95 (95% confidence interval: 3.59–4.31), on a 5-point scale running from 1 (*not at all*) to 5 (*a lot*). Participants also reported to have noticed the seatbelt very well (mean = 4.6 on a scale running from 1 (*poorly noticed*) to 5 (*strongly noticed*), 95% confidence interval 4.32–4.88).

5.2. Stiff brake pedal (longitudinal deceleration cueing)

Considerable effects were observed for the stiff brake pedal. Stopping consistency improved, whereas maximum decelerations and onset jerk were lower for the stiff pedal as compared to the soft pedal. The questionnaire showed that participants who drove with the stiff pedal gave a higher rating to a question that asked how realistic the braking pedal felt to them on an ordinal scale from 1 (*poor*) to 5 (*very well*) (mean Soft = 2.55, mean Stiff = 3.05, Cohen's *d* = -0.47, *p* = 0.096 using a paired *t* test, or *p* = 0.012 using a nonparametric sign test).

5.3. Vibrating steering wheel (longitudinal deceleration cueing)

The vibrating steering wheel had a relatively small effect on braking behaviour. The only significant effect was that vibrations On resulted in a lower onset jerk than Off.

5.4a. Screeching tyre sound (longitudinal deceleration cueing)

The screeching tyres showed no significant effects in the means of the dependent measures. However, an important aspect of the screeching sound is that it can act

Table 4. Results of the experiments

	1. Seatbelt system				2. Brake pedal			
	Off	On	*p*	*d*	Soft	Stiff	*p*	*d*
NrValidStops	9.25	8.65	0.07	0.42	9.17	9.29	0.7	-0.11
Mean Vini	17.2	16.2	0.007	0.43	15.6	15.2	0.08	0.47
Mean DTLini	62.2	79.1	0.004	-0.83	41.1	45.1	0.2	-0.25
SD DTLfin	3.45	2.51	0.006	0.63	3.86	2.59	0.002	0.75
Mean max. dec.	6.27	4.57	<0.001	0.85	7.65	6.53	<0.001	0.69
Mean onset jerk	5.73	2.30	<0.001	0.99	12.99	6.88	<0.001	0.96

	3. Vibrating wheel				4a. Screeching tyres			
	Off	On	*p*	*d*	Off	On	*p*	*d*
NrValidStops	9.69	9.69	1.0	0.00	7.13	6.61	0.3	0.29
Mean Vini	15.5	15.0	0.2	0.25	13.8	14.1	0.5	-0.18
Mean DTLini	39.3	40.2	0.7	-0.10	40.0	36.9	0.3	0.23
SD DTLfin	2.01	1.94	0.8	0.07	2.33	1.73	0.3	0.37
Mean max. dec.	6.88	6.57	0.2	0.26	6.61	6.47	0.7	0.09
Mean onset jerk	7.53	5.35	0.023	0.56	6.03	5.58	0.7	0.10

	4b. Screeching tyres				5. Vibrating seat			
	Off	On	*p*	*d*	Off	On	*p*	*d*
NrDepartures	0	0	1.0	0.00	0.13	0.07	0.3	0.22
Mean LCE45deg (L)	1.44	1.15	0.4	0.19	1.61	1.60	1.0	0.01
Mean LCE45deg (R)	-0.88	-0.79	0.3	-0.20	-0.88	-0.83	0.8	-0.07
Mean V45deg (L)	9.41	8.95	0.2	0.33	9.12	9.14	0.9	-0.02
Mean V45deg (R)	8.54	8.55	1.0	0.00	8.07	7.29	0.044	0.50

	6. Pressure seat			
	Off	On	*p*	*d*
NrDepartures	1.62	1.22	0.5	0.25
Mean LCE45deg (L)	1.55	1.25	0.3	0.38
Mean LCE45deg (R)	-1.14	-0.78	0.1	-0.59
Mean V45deg (L)	10.8	10.5	0.5	0.21
Mean V45deg (R)	10.6	9.93	0.4	0.36

Note. *p* represents the probability of observing the given means by chance if the means would actually be equal. *p* was calculated using a *t* test.
d represents Cohen's *d*, that is, the difference between the means divided by the pooled standard deviation.

like a binary warning signal, leading to behavioural adaptation. Indeed, the mean maximum deceleration was 7.1 m/s^2 (SD = 1.2) in runs 1 and 2, which significantly decreased to 5.9 m/s^2 (SD = 1.6) in runs 3 and 4 ($p = 0.002$), a learning effect. An additional control group ($n = 5$, not shown in Table 1) driving without any screeching tyre sound showed constant maximum deceleration over time: 6.3 m/s^2 (SD = 1.9) in runs 1 and 2, and 6.2 m/s^2 (SD = 2.5) in runs 3 and 4. So, participants decreased decelerations with run number, whereas this effect was not found for a control group who drove entirely without screeching.

5.4b. Screeching tyre sound (lateral acceleration cueing)

There were no significant effects in the means of the dependent measures. As with the longitudinal experiment (section 5.4a), there were indications that participants responded to the screeching tyres: when selecting only the first turn from each run, the speed with On was lower than with Off in left and right turns ($p = 0.027$ for the left turns). Results of a questionnaire indicated that the majority of participants who had heard the screeching tyres indicated that it positively influenced their driving behaviour, mainly by adopting lower cornering speeds.

5.5. Vibrating seat (lateral acceleration cueing)

The speed halfway in right turns was significantly lower with On as compared to Off. The distance to the road side was not significantly different between On and Off. Plots of the mean vehicle paths for On and Off revealed that these were highly similar throughout the turn (data not shown). No measure was significantly different between On and Off for left turns.

5.6. Pressure seat (lateral acceleration cueing)

There were no significant differences between On and Off for any of the measures. An important finding was that some participants indicated afterwards that they did not understand the purpose of the pressure seat. That is, they had not understood that the pressure, which they felt clearly on their back, corresponded to the lateral acceleration (cornering speed) of the car. Analyses of the questionnaire revealed that respondents who drove with On tended to give a lower score to an item that asked to judge to what extent they had the feeling of driving in a real car (mean Off = 3.00, mean On = 2.56, Cohen's $d = 0.53$, $p = 0.16$, on a 5-point scale running from 1 (*strongly disagree*) to 5 (*strongly agree*).

6. Discussion

All systems, except for the pressure seat, showed desired effects, that is, braking and cornering behaviour became more like real-life driving behaviour. The magnitude and the nature of the effects, however, were notably different between the systems.

The seatbelt tensioning system had large positive effects on braking behaviour, which is remarkable when considering that the cue is not realistic by itself. A seatbelt in a real car does not produce a force: It allows freedom of movement by the driver, unless a locking mechanism causes that seatbelt forces can rise during sudden vehicle decelerations. In addition, the seatbelt tensioning system improved subjective realism of the simulator. This result corresponds to earlier studies using a motion seat in a fixed-base simulator that found that participants rated the simulator with a motion seat as more realistic, irrespective of what kind of motion cueing algorithm was applied (Mollenhauer et al., 2004; chapter 7). This gives the impression that an illusion of reality can relatively easily be provided to a driver in a simulator. It remains to be seen, however, whether the seatbelt improves training effectiveness, as problems may occur when transferring to real-life situations in which the cue is absent.

A striking result is that a simple modification to the brake pedal had large effects on braking behaviour. It is recommended to further investigate the effects of changes to the force-feel characteristics of the brake system. According to chapter 6, a realistic (brake) pedal addresses a task-relevant cue, can lead to positive reactions, and more realistic driving behaviour without inducing negative effects that are associated with increased levels of fidelity such as distraction, high costs, or complicating factors.

Stopping consistency improved for the stiffer pedal as well as for the seatbelt tensioning system. A similar result was found for a motion platform (Siegler et al., 2001) and a motion seat (chapter 7). It seems that providing additional deceleration cues improves the ability to precisely stop a vehicle, irrespective of *how* these cues are presented.

Screeching tyres were a simple modification that led to behavioural adaptation. A caveat is that only a driver who drives (very) fast gets to hear the screeching tyre sound, whereas a slower driver does not. Because many driving simulation studies report on unrealistically fast driving, screeching sound seems a justified means to deal with this undesired behaviour.

Steering wheel and seat-based vibrations also produced desired effects but of relatively small size. The intensity of the vibrations, however, were rather weak in the present experiments. Many participants did not notice the vibrating steering wheel. Therefore, it is recommended to use vibrations of sufficient intensity.

The pressure seat was hardly successful in lowering cornering speeds, although the effects were in the desired direction. It was worrying that some participants failed to notice that the stimulus was related to vehicle speed. A similar result was found by Showalter and Parris (1980), who found that a g-seat providing acceleration cues to pilots in a simulator led to little, if any, performance improvements.

The present results as well as previous research using a motion seat (chapter 7), suggest that it is difficult to influence lateral accelerations using nonvisual motion cues. A likely explanation is that a driver has a good notion of lateral acceleration in turns because his or her speed can be perceived visually. During straight line driving (e.g., braking), however, the driver cannot infer vehicle acceleration information directly through the optic flow (this requires a differentiation), and therefore has to rely more on nonvisual cues. It is noted that comparisons between experiments should be made with a certain reticence; statistical power varied as a function of group size and experimental design. Additionally, it would be interesting to determine the effects of combinations of feedback systems used concurrently.

It is concluded that diverse motion cueing solutions can be successful in improving absolute fidelity in a fixed-base simulator. The best results were obtained with stimuli of sufficient intensity and with good stimulus-response compatibility. It is recommended that researchers and simulator developers first implement the simplest systems for improving absolute fidelity, such as a correct brake pedal stiffness, because these may already have profound effects. When such aspects have been properly implemented and validated, one could refer to more expensive and more complex solutions such as a motion seat.

CHAPTER 9

Feedback on mirror-checking during simulation-based driver training

Abstract

Current driver-training simulators do not provide feedback on students' mirror-checking (MC) performance, which could be detrimental to training effectiveness. This chapter studies the effect of feedback on learning the MC task. After completing a short training session to become acquainted with the simulator, two groups of inexperienced drivers and two groups of experienced drivers (*n* = 10 for each group) completed a simulator lesson on MC. Using webcams, MC performance was assessed by a human experimenter who was unaware as to which group the participants belonged. Two groups were provided with sampled feedback on their MC performance; the others were not. Retention was assessed during a subsequent lesson on cornering behaviour during which no feedback on MC was provided to any of the groups. Results showed that, during the cornering lesson, the experienced feedback group performed significantly better in MC than the experienced no-feedback group. No statistically significant learning effect was found for the inexperienced groups. It is concluded that feedback on MC was useful for experienced drivers. Optimization of feedback and instruction is expected to lead to further gains in simulator training effectiveness.

De Winter, J.C.F., De Groot, S., Van Loenhout, M.J., Van Leeuwen, A., Do, P., Wieringa, P.A., & Mulder, M. (2008d). Feedback on mirror-checking during simulation-based driver training. *Proceedings of the 27th European Annual Conference on Human Decision Making and Manual Control*, Delft, the Netherlands. (adapted with minor textual changes)

1. Introduction

Every year, a significant number of road crashes occur as a result of inappropriate checking of the rear view mirrors and failures to look at blind spots. Moreover, incorrect viewing behaviour belongs to the most frequently occurring errors during the official Dutch driving test (see Table 1). For these reasons, it is important that viewing behaviour – mirror-checking (MC) in particular – is taught well during driver training.

The use of cost-effective driving simulators with intelligent tutoring systems is on the rise. Such training simulators have the capability to automatically instruct a student to perform an MC sequence. Moreover, the use of mirrors can be trained by showing the need for MC, such as by including events with adjacent vehicles in the driver's blind spots. However, current simulators lack the ability to assess whether someone has indeed looked into the mirrors and therefore cannot provide feedback on MC (Kappé & Van Emmerik, 2005).

Several techniques are on the market that can measure a person's eye or gaze direction without human intervention. These have been successfully applied in scientific research. However, depending on technology, eye-tracking can be intrusive if calibration is necessary, or if additional equipment should be worn by a student. Moreover, difficulties may occur with respect to individual posture, head movement, or eyeglasses. Finally, eye-tracking can be rather expensive, sometimes as expensive as an entire simulator. Although low-cost solutions are under development (Fikkert et al., 2006), eye-tracking is not yet used in simulation-based driver training.

Even if the technical difficulties of eye-tracking are dealt with, it remains uncertain whether an eye-tracking device can improve the effectiveness of simulation-based driver training. Surely it is possible to incorporate eye-tracking in a simulator in order

Table 1. The 10 most recorded errors (out of 150 possible) amongst a selected group of 370 candidates who failed on the first attempt of the official Dutch driving test

1.	Viewing behaviour on/near intersections	58%
2.	Viewing behaviour while changing direction	42%
3.	Position on the road while changing direction	31%
4.	Viewing behaviour while overtaking	30%
5.	Right of way on/near intersections	27%
6.	Adaptive/decisive driving on/near intersections	25%
7.	Position on the road while overtaking	23%
8.	Viewing behaviour while merging entry/exit lanes	20%
9.	Position on the road while driving on straight/curved road segments	19%
10.	Viewing behaviour while changing lanes/lateral movements	18%

Note. The mean number of errors per candidate was 6.13. The data were obtained from the Dutch Driving Test Organization and were from the period August 2005 – March 2006.

to provide automatic feedback on MC, but how important is this technology for training effectiveness? Although Kappé and Van Emmerik (2005) stated that the current lack of eye-tracking measurements is a vital drawback of simulation-based driver training and testing of drivers, no objective data have been presented.

The aim of this study was to investigate the learning effects of feedback on MC performance during a typical initial driver training lesson of negotiating intersections. Although the main target of driving training simulators is to tutor inexperienced drivers, simulators can also be used as a means to provide fresh-up courses for experienced drivers. These drivers should already be familiar with the mirror task on the roads but may have abandoned the correct procedure. Hence, the effects of feedback on MC were investigated for both inexperienced and experienced drivers.

2. Method

2.1. Participants

Four independent groups of 10 participants each were formed (see Table 2), all recruited from the Delft University of Technology. Two groups consisted of drivers without any on-road experience (I groups); the other two groups consisted of experienced drivers (E groups). An experienced driver was defined as someone who had a driving licence or someone who had on-road experience in the form of lessons with a professional driving instructor. Two groups received feedback on MC (the I-F and E-F groups); the two other groups did not receive feedback (the I-NF and E-NF groups). The two E groups as well as the two I groups were matched with

Table 2. Composition of the four groups

Group [a]	Mean age	Men/ women	Driving licence [b]	Driving lessons [c]	Simulator [d]	Moped [e]	Go-kart [f]
I-NF	18.5	7/3	0	0	1	2	4
I-F	18.0	7/3	0	0	0	1	5
E-NF	20.3	9/1	5 (4.1)	5 (15.8)	3	4	7
E-F	20.4	9/1	7 (3.1)	3 (18.8)	1	4	7

[a] I-NF: inexperienced drivers, no feedback. I-F: inexperienced drivers, feedback. E-NF: experienced drivers, no feedback. E-F: experienced drivers, feedback.

[b] Number of participants with licence (mean number of years of licence possession).

[c] Number of unlicensed participants with driving experience in the form of driving lessons from a professional instructor (mean number of hours).

[d] Number of participants who had experience in a driving simulator.

[e] Number of participants who had experience with a moped with a manual clutch.

[f] Number of participants who ever drove a go-kart.

Table 3. Simulator driving lessons that participants completed

Lesson 1 (L1). Driving away (identical for each group). Multimedia instruction, 3.5 min training, questionnaire.
Lesson 2 (L2). Mirror checking (feedback on mirror-checking for two of the four groups). Multimedia instruction, 8 min training (at least 14 intersections), questionnaire.
Lesson 3 (L3). Cornering (identical for each group, no feedback on mirror-checking). Multimedia instruction, 8 min training (at least 14 intersections), questionnaire.

respect to age, gender, and amount of experience. Participants were not informed about the purpose of the experiment.

2.2. Apparatus and procedures

Participants completed three lessons (summarized in Table 3) in a fixed-base Dutch Driving Simulator (Green Dino, 2007). The entire experiment, including the automated multimedia instructions and questionnaires, lasted about 30 min per participant.

To become acquainted with the simulator, participants first engaged in an automated drive-away lesson (L1) with virtual instruction, which lasted 3.5 min. This lesson took place on an endless straight road and was identical for all groups.

Next, participants completed a lesson on MC (L2). After an automated multimedia instruction on how to check the mirrors before changing direction, the lesson started. Participants only had to steer; speed-control, starting the engine, and using the parking brake were automated. The lesson took place in an environment that primarily consisted of unsignalized intersections. No other traffic was present. Upon approaching each intersection, participants were instructed to turn left or right. Upon approaching the 2nd, 3rd, 5th, 7th, 10th, and 13th intersections, participants of all groups received preprogrammed verbal instructions to check the mirrors (see Table 4). After each intersection, participants belonging to the F groups received feedback on their MC performance (see Table 5). Participants in the NF groups did not receive such feedback.

Finally, to test the retention of the learning of MC, all participants completed a cornering lesson (L3). After an automated instruction on how to approach corners, the lesson started in the same virtual environment as L2. Participants had to steer, use the throttle, and brake themselves, raising the task difficulty compared to the

Table 4. Possible instructions during Lesson 2 and Lesson 3 (translated from Dutch)

Turn left/right
Look into the inside mirror, the outside mirror, and to the side
Mind your speed in turns [a]

[a] This instruction was only used during Lesson 3.

Table 5. Possible feedback during Lesson 2 (translated from Dutch)

You have looked well
You haven't looked into the inside mirror
You haven't looked into the outside mirror
You haven't looked to the side
You haven't looked well into the mirrors [a]

[a] This feedback was provided when a participant made more than one mistake in a mirror-checking sequence.

previous lesson and directing participants' attention away from the MC task. Upon approaching each intersection, participants were instructed to turn left or right. Upon approaching the 2nd, 7th, and 10th intersection, participants received instructions to check the mirrors. None of the groups received feedback on MC.

After each lesson, participants had to fill in a questionnaire on their subjective experience (see section 2.4.3).

2.3. Assessment of mirror-checking

During L2 and L3, two experimenters monitored the participant's behaviour using three webcams that were placed in the simulator. The webcams did not disturb the participant's view on the road. The two experimenters were each sitting in front of a computer outside the view of the participant and independently assessed whether the participant had correctly looked into the mirrors and to the side. A computer programme made use of the assessments of one of the experimenters for providing presampled verbal feedback (see Table 5). The assessments of the second experimenter were not used during the experiment, but only after all data were recorded, for making interexperimenter comparisons. The software on the experimenters' computers was programmed in such a way that the experimenters could not distinguish as to which group participants belonged. Because participants wore headphones, the experimenters could not hear whether the participant received feedback or not.

2.4. Independent and dependent variables

A between-subjects design was employed. The independent variable under investigation was whether feedback on MC was applied during L2 or not. We compared the I-NF versus I-F group, and E-NF versus E-F group. The dependent measures are described in sections 2.4.1 through 2.4.3.

2.4.1. Mirror-checking score

Before each intersection, participants had to check two mirrors and look over their shoulder at the blind spot. Accordingly, 3 score points could be obtained per inter-

Table 6. Objective performance measures (Lesson 2 and Lesson 3)

Speed [m/s]	Mean speed during the lesson
SDLP [m]	Standard deviation of lateral position on straight road segments
Indicator [%]	Percentage of intersections approached with indicator *on*

section. The total MC score of the first 14 valid intersections was calculated for each participant (i.e., maximum 3 x 14 = 42 points per lesson) and linearly scaled from 0 to 100%. Intersections on which participants turned into a direction other than the instructed direction were declared invalid. In case of missing values or inter-experimenter disagreement, the stored videos were used to come to an agreement.

2.4.2. Objective measures

Table 6 shows the objective measures that were used. Earlier research has shown that speed and lane keeping accuracy (standard deviation of lateral position or SDLP) are sensitive measures (De Winter et al., 2006d). A lower mean speed or a higher standard SDLP could be indicative of increased workload. Participants were not told to use the turn indicator, except briefly as part of the multimedia introduction of L2. Using the direction-indicator lights can therefore be considered an embedded secondary task. Earlier research has shown that indicator errors are a sensitive measure to changes in task load demands (De Groot et al., 2006).

2.4.3. Subjective measures

After each lesson, participants had to fill in a questionnaire comprising 10-point scales running from 1 (*No*) to 10 (*Yes*). Of specific interest were the questions shown in Table 7, as these concerned the face-validity of the MC learning and aspects related to participants' overall satisfaction. Other queries (not shown) were regarded as "dummies", included to direct participants' attention away from the fact that the research focus in this experiment was on MC.

Group differences were evaluated using an independent *t* test. If the *p*-value was below the 0.05 threshold, then the null hypothesis of equal means was rejected.

Table 7. Questionnaire items (translated from Dutch)

Q1	Do you think you have well learned the use of mirrors?
Q2	Do you think you have well performed the use of mirrors?
Q3	Do you think that the basic principles of MC can be well taught using a simulator?
Q4	Did you enjoy driving in the simulator?
Q5	During the cornering lesson, did you still think about the mirroring lesson?
Q6	Did you find the simulator realistic?

Note. Q1–Q3 were posed after Lesson 2; Q4–Q6 were posed after Lesson 3.

Table 8. Mean mirror-checking scores

	Inexperienced drivers				Experienced drivers			
	I-NF	I-F	*d*	*p*	E-NF	E-F	*d*	*p*
Lesson 2	72	88	-1.26	0.014	88	97	-1.03	0.039
Lesson 3	64	67	-0.10	0.8	75	95	-0.84	0.039

Note. *p* was calculated using a *t* test. *d* represents Cohen's *d*, that is, the difference between the means divided by the pooled standard deviation.

Cohen's *d* was used as an effect size measure. It is defined as the difference between the means divided by the pooled standard deviation.

3. Results

3.1. Mirror-checking score

Table 8 and Figure 1 show the MC scores. During L2, the F groups performed significantly better than the NF groups. However, feedback resulted in higher retention of MC performance during L3 for the E-F group only.

3.2. Objective measures

Table 9 shows the results. Although trends can be recognized, no statistically significant effects between NF and F were found.

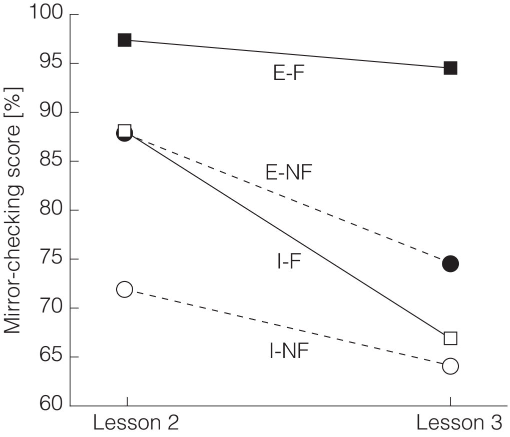


Figure 1. Mean mirror-checking scores.

Table 9. Means of the objective measures

	Inexperienced drivers				Experienced drivers			
	I-NF	I-F	*d*	*p*	E-NF	E-F	*d*	*p*
Speed (L2) [m/s] [a]	13.2	13.3	-0.09	0.9	12.9	13.4	-0.42	0.4
Speed (L3) [m/s]	13.5	13.1	0.18	0.7	14.6	13.6	0.53	0.3
SDLP (L2) [m]	0.37	0.43	-0.47	0.3	0.29	0.27	0.28	0.5
SDLP (L3) [m]	0.38	0.43	-0.43	0.3	0.28	0.27	0.17	0.7
Indicator (L2) [%]	30	66	0.78	0.098	80	93	0.39	0.4
Indicator (L3) [%]	34	53	0.41	0.4	87	97	0.47	0.3

[a] Speed control was automated and therefore subject to little variability only.

3.3. Subjective measures

Results are shown in Table 10. The E-F group had indicated that they learned less well to check the mirrors than the E-NF group (Q1). Moreover, the I-F group found the simulator significantly less realistic than the I-NF group (Q6).

3.4. Additional observations

Table 11 shows comparisons of the experimenters' MC assessments. It can be seen that the number of disagreements was small, indicating that the MC was reliably assessed.

During L2, some participants always looked to the left, independent of whether turning left or right on the intersections. Apparently, they had followed the multimedia instructions literally, that is, not within the context of the intersection event. Therefore, in-between L2 and L3, an experimenter gave these participants a brief verbal explanation on how they should check the mirrors. In total, 4 participants from the I-

Table 10. Means of the subjective measures

	Inexperienced drivers				Experienced drivers			
	I-NF	I-F	*d*	*p*	E-NF [a]	E-F [b]	*d*	*p*
Q1	6.5	7.6	-0.68	0.1	8.1	6.9	1.14	0.034
Q2	6.7	7.8	-0.59	0.2	8.6	8.6	-0.02	1.0
Q3	8.5	7.4	0.73	0.1	7.9	8.0	-0.07	0.9
Q4	8.1	7.8	0.25	0.6	8.4	8.1	0.24	0.6
Q5	7.9	8.3	-0.22	0.6	8.3	7.6	0.45	0.4
Q6	7.2	5.4	0.99	0.041	5.0	5.9	-0.41	0.4

[a] 1 participant had not responded.
[b] 3 participants had not responded.

Table 11. Interexperimenter comparisons for the 1,059 available MC sequences

	1. Inside mirror	2. Outside mirror	3. To the side
Pass/Pass	88.7%	84.1%	69.3%
Pass/Fail	1.0%	1.3%	1.9%
Fail/Pass	1.0%	0.9%	2.3%
Fail/Fail	9.3%	13.6%	26.5%
Total	100%	100%	100%

Note. Each mirror-checking sequence consisted of three actions.

NF group, and 2 from the I-F group received these instructions, a difference which was not statistically significant (p = 0.4 using a Wilcoxon test).

4. Discussion

This study showed that feedback on mirror-checking (MC) improved MC performance, which is one of the most important driving tasks. However, a beneficial effect of feedback on retention was found only for the experienced drivers.

The E-F group indicated in the questionnaire that they learned *less* well to check mirrors than the E-NF group. A possible explanation is that feedback was interpreted as an annoying disturbance for the experienced drivers who already knew how to check the mirrors. The I-F group believed that the simulator was *less* realistic than the I-NF group. A possible explanation is that feedback was not perfectly consistent in terms of timing, which led to the belief that the entire simulator was less realistic. Another explanation is that inexperienced drivers misconceived the meaning of realism, as they had never driven a real car before.

Several possible reasons for the insignificant learning effect of feedback amongst inexperienced groups can be thought of. The primary task of vehicle control may have been too demanding so that participants failed to learn the MC task. It is also possible that there were indeed differences, but the data were not sensitive enough to detect them. The sample size in this study can be considered reasonable, however.

As indicated in section 3.4, some of the inexperienced participants initially made errors because they had misunderstood how to check the mirrors. Clearly, proper instructions are crucial for training effectiveness. It is recommended to first optimize such evident didactic aspects, including the position of the lesson in the overall curriculum, before spending considerable resources to eye-tracking technology. Recommendations on how to improve didactics and hardware characteristics in simulation-based training can be found in De Groot et al. (2007) and chapter 6. Several other limitations apply. Webcam stagnation sporadically occurred. These were random events, however; there were no indications that they interacted with the experimental conditions. The participants were recruited from the university community, which has particular sociodemographic characteristics. Moreover, participants' motivation during an experiment may be different than during an actual simulator lesson. Finally, it is possible that participants became aware of the aim of the study during the course of the experiment. These factors could delimit the generalizability of the present work.

In the present study, we provided simple feedback on MC. Eye-tracking devices provide supplementary possibilities, for instance, presenting students with a two-dimensional "map" of their visual scanning behaviour. Moreover, when subjecting drivers to a standardized simulation-based test, aberrant driving behaviour, such as incorrect MC, can be objectively identified (Chapman et al., 2002). This is likely to be an advantage of eye-tracking compared to on-road driving during which performance is subjectively assessed under nonstandardized conditions. Besides, scientific reviews have indicated that formal on-road driver training is not safety-effective (e.g., Elvik & Vaa, 2004). Therefore, we believe it is out of place to characterize the lack of eye-tracking as a disadvantage of driving simulation. Instead, eye-tracking can be regarded as a relatively new and unexplored opportunity for improving driver training and testing.

CHAPTER 10

Conclusion, discussion, and recommendations

This chapter provides the main conclusions, elaborates on the results of this thesis, outlines its limitations, and offers recommendations for future studies in the field of driving simulation.

1. Conclusion

The first objective of this thesis was to develop a method that can process data obtained from driving simulators into a valid student-profile on driving skill and driving style. The second objective was to provide a fundamental understanding of the relationships between driving simulator fidelity and training effectiveness, and specifically to investigate whether low-cost motion cueing systems are valuable in simulation-based driver training.

The fist conclusion is that the statistical method *exploratory factor analysis* is suitable for processing large amounts of data into a smaller number of theoretically sound personal scores. The extracted speed-, violation-, and error-scores were meaningful individual measures, as indicated by their correlations with gender, age, and driving test results.

The second conclusion is that fidelity requirements depend on a compromise, in which negative effects, such as costs, simulator sickness, and distraction, have to be weighted against positive effects, such as improved validity and user acceptance. Low-cost motion cueing simulators are able to increase subjective ratings of realism and improve in-simulator driving performance as compared to control conditions without the motion system. However, validation research is still needed with respect to transfer of training to the roads.

2. Discussion

2.1. Introducing factor analysis for driver assessment

Based on a data analysis of a driving simulator experiment and by means of a literature overview on driver behaviour models (chapter 2), the value of motivational models (in particular the risk homeostasis theory (RHT) and the task capability interface (TCI)) was investigated. The overview illustrated that due to the lack of specificity and lack of clear boundary conditions, it is possible to find an explanation for virtually *any* experimental result. In other words, the RHT and TCI cannot be falsified, failing to fulfil one of the most important criteria of scientific models. It was concluded that motivational models – although useful from a conceptual viewpoint – cannot be used for constructing a quantitative student-profile from simulator data.

Adaptive control models suffer from the opposite problem. These types of models have been very useful for quantitative driver parameter identification in basic tasks, such as car following, curve following, and regulation against wind gusts. Measures could include crossover frequency, time delay, and coherence (and these measures can still be included in a factor analysis). However, there are a number of pitfalls. First, it is always possible to construct a model and to subsequently fit the data to it in order to account for a certain share of the variance. However, this does not mean that the model is necessarily the best or even an adequate model. The transfer functions (blocks) and information flows (lines) are inevitably arbitrarily chosen. Chapter 2 provided some examples showing that strong latent correlations can be present in the data which are not readily recognized or understood by adaptive control models. Moreover, certain adaptive control models have been criticized for their psychological implausibility and their failure to adequately incorporate individual differences. In addition, we discussed the critical problem of overfitting and the lack of validation. We recognized the tendency in adaptive control models towards increased model complexity in order to provide a good fit to experimental data. This good fit, however, comes at a price of parameters that have limited generalizability, or parameters that may be meaningless.

Understanding driving – a human-machine process – requires a *statistical model* that explains individual differences, instead of a deterministic model that tries to explain exactly what a driver does at a certain moment. Specifically, we propose exploratory factor analysis (EFA) as a technique for constructing a student-profile. EFA deviates from methods that lock driver assessment into single variables and causal relations. Instead, EFA *uses* the multivariate correlation structure to retrieve underlying patterns and to explain drivers' behaviour parsimoniously. EFA can be interpreted as a method that exploits the *principle of the common cause*, facilitating the generation of plausible scientific theories (Haig, 2005).

The use of EFA for studying individual differences is neither new nor unexpected. Factor analysis is one of the most widely used techniques in psychological research (Fabrigar et al., 1999; Cudeck & MacCallum, 2007) and is increasingly recognized in other scientific domains (Kaplunovsky, 2005). The method has been broadly applied in driver behaviour studies as well, mostly based on self-reports. A well-known example is the Driver Behaviour Questionnaire (DBQ) which has been used for identifying violations and errors (and more specific factors such as aggressive violations, interpersonal violations, slips, and lapses; Reason et al., 1990). Although self-reports are able to recover valuable information, it can be problematic to rely solely on this type of data. For example, Bjørnskau and Sagberg (2005) found different results as a function of driver experience when an ordinal DBQ was used that asked *how many times during the last month* a certain behaviour was committed, instead of the traditional 6-point scale ranging from 1 (*never*) to 6 (*very often*). Ironically, asking drivers how many times they had slips or lapses is problematic in itself: "Unconscious errors may be hard to remember precisely because they are unconscious; they are not something we want or plan to do" (Bjørnskau & Sagberg, 2005, p. 137).

The present thesis factor analysed learner drivers' data from a diverse range of tasks in standardized virtual environments (chapters 3 and 4). Analyses of large pools of trainees' performance records seem scarce in the scientific literature. Exceptions are Allen et al. (2006; 2007a) and Turpin et al. (2007), but these have not reported on a factor analytical approach.

2.2. Violations, errors, and speed factor-score predictors

Chapter 2 provided an example of the use of EFA for driver assessment. A speed factor was extracted from a series of performance measures on an intersection encounter task. Results showed that the speed-score reliably correlated with gender and the interpersonal violation-score of the DBQ, demonstrating that the speed-scores conveyed meaningful information about driver behaviour. Various measures (e.g., minimum speed, safety indicators) had high loadings on the same speed factor and therefore essentially represented the same construct. A possible limitation of this study was that the performance measures were drawn from the same experiment and were not experimentally independent.

In chapter 3, we extended the use of EFA by employing the method on more diverse tasks of a pool of 520 students who completed an initial driver-training programme in a simulator. According to the well-known violation-error distinction originally proposed by Reason et al. (1990), two factors were extracted from the students' task failures: violations and errors. Most likely, this was the first time that the violation-error distinction has been retrieved from objective driver data. The distinction between errors and violations seems to correspond to other constructs in traffic psychology, such as driver performance and behaviour (Evans, 2004), driving skills and style (Elander et al., 1993), and skills and safety motives (Lajunen & Summala, 1995).

Next, we used EFA on students' mean task completion times. Despite the diversity of the driving tasks, it was clear that one factor provided the best explanation. The extracted factor was again interpreted as a speed factor (paceability factor in chapter 3), and the factor loadings were interpreted as the extent to which students' inclination for quick task-execution was expressed in the task completion times. S*elf-paced* tasks loaded positively on the speed factor, *forced-paced* tasks had approximately zero loadings, and *inverse-paced* tasks had negative factor loadings. Tasks of the latter category actually took longer to complete for the quicker drivers.

Chapter 4 used EFA to extract a violation factor, an error factor (called steering-errors factor in chapter 4), and a speed factor, now based on a larger sample of participants and using more elementary driving tasks in more training sessions. The predictive validity of the factor scores was investigated by calculating correlations with the results of the on-road driving test. Earlier licensure was statistically associated with a lower violation-score, a lower error-score, and a higher speed-score.

Participants with a higher speed-score completed more tasks, committed more violations, and fewer errors. In addition, it was shown that errors and violations were slightly negatively correlated, indicating that these are two different forms of aberration, in agreement with the original hypothesis of Reason et al. (1990).

Chapters 2–4 showed that there were large gender differences, in some cases larger than one standard deviation. Men had a higher speed-score, a higher violation-score, and a lower error-score than women. This is in accordance with DBQ-results, in which men on average have a higher violation-score but a lower error score (lapse score) than women (Reason et al., 1990). The gender differences seem in good agreement with gender differences in on-road crash involvement (Clarke et al., 2005; Laapotti & Keskinen, 2004).

Correlations between factor scores and driver age were calculated as well. These correlations had a magnitude that was smaller than 0.2 (chapter 4). The small magnitude of the correlations can be explained by the notion that the age range was rather small: virtually all simulator drivers were younger than 30 years. When calculating correlations between age and driving speed/performance in the simulator for participants across a wider age span, correlations were stronger in magnitude, up to 0.6 (De Winter et al., 2006a; H.C. Lee, 2003).

It is well established that age has a positive effect on road safety (chapter 1; OECD, 2006). Serious accidents, and in particular fatal accidents, are often the result of extreme forms of behaviour, such as excessive speeding. Figure 1 shows age and gender differences for the more extreme kinds of speeds and violations (i.e., a score greater than 1) in the simulator. The figure shows that these extremes were particularly prominent amongst young men. Note that the effects of age are not necessarily causal effects of biological age but could be confounded through age-related volunteer bias to engage in simulator training (see appendix A for an elaboration on this matter).

2.3. Quantifying violations

Literature suggests that violations and errors are mediated by different psychological mechanisms (Parker, 2007; Reason et al., 1990). Violations are – at least partly – a result of voluntary behaviour (i.e., what a driver chooses to do), whereas errors can be best explained by cognitive processing limitations (i.e., what a driver is able to do). Quantifying a drivers' tendency for committing violations is important because violating behaviour relates to crash involvement (Parker et al., 1995; Parker, 2007) and “adequate psychomotor skills and physiological functions are not sufficient for good and safe performance as a driver” (Hatakka et al., 2002, p. 202).

One may argue that students do not commit violations during driving lessons. Indeed, simulators have been regarded as devices that can be only used for assessing driving skills (e.g., reaction times, control accuracy) rather than driving style (Evans, 2004). We infer, to the contrary, that students engaged in prolonged periods of self-paced driving in a forgiving environment, independent of human instructors, will inevitably reveal their “true nature”. Inevitably, some drivers choose to accept a gap with less hesitation than others, are more inclined to break the speeding code when the situation is considered relatively safe, or are less inclined to stop in front of a traffic light when there is just enough time to stop. Violations can become part of a routine when somebody believes to have sufficient skills to violate the rules habitu-

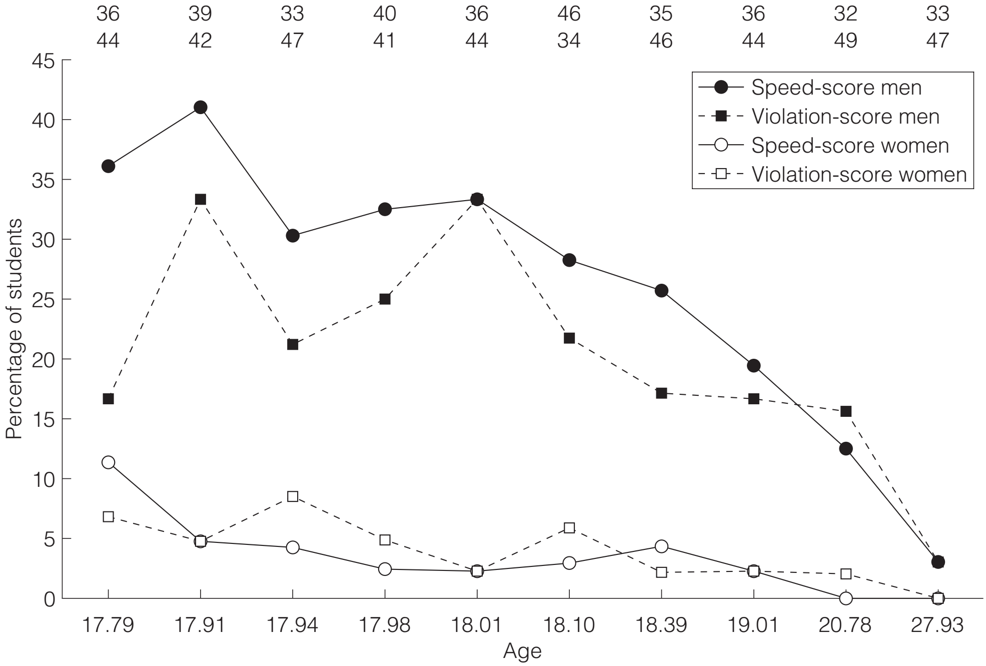


Figure 1. Percentage of men/women with a speed-score/violation-score greater than 1. This figure was created using data of the study group of chapter 4 (*N* = 804). Participants were sorted according to their age at the first driving simulator lesson and subsequently divided into 10 groups of 80 or 81 each. The horizontal axis represents the mean age of the 80/81 participants in the groups. The vertical axis represents the percentage of men/women. The numbers in the top of the figure represent the number of men per group (top row) and number of women per group (bottom row). The error-score had a more erratic pattern as a function of age and is therefore not shown. 4.6% of the men and 14.6% of the women had an error-score greater than 1.

ally (Parker, 2007). Committing violations has also been interpreted as the crossing of a barrier: Whether or not a barrier will be crossed depends on the costs and benefits associated with the barrier-crossing (Nap et al., 2007; Polet et al., 2002). As indicated by Reason et al. (1990), violations are not necessarily reprehensible. In some cases, the rules can be too strict compared to the situation, leading to invention of the term *correct violations* (Reason, 1999).

A number of studies have also recognized that simulators are able to capture motivational components of car driving, not only driver skills. It has been found that sensation seeking scores (Schwebel et al., 2006), self-reported driver aggression (Deffenbacher et al., 2003; Matthews et al., 1998), self-reported driving style (Hoedemaeker & Brookhuis, 1998), and at-risk personality characteristics (Deery & Fildes, 1999) were predictive of driver behaviour in a simulator. Driving simulators have also been used recently to study aggressive driving and road rage (Drews et al., 2003), running yellow/red lights (Allen et al., 2005b; Senserrick et al., 2007), voluntary violations (Chalmé et al., 2006), and violating behaviour under time pressure (Bonsall & Palmer, 1997). An interesting finding of Chalmé et al. (2006) was that there were large individual differences: There were participants who (almost) never voluntarily violated a rule and others who made a large number of violations.

Summarizing, a driving simulator can capture more than just skills: A driving simulator can capture motivational factors including violations and choice of speed. The present thesis provided a method that can identify violation-prone students before they head to the roads, which could be highly relevant to road safety. However, more validation research is needed before affirmative conclusions can be made in this respect (see section 3.5 as well).

2.4. The speed factor

The existence of a speed factor is consistent with motivational models of driver behaviour in which speed is generally regarded as the primary variable (Vaa, 2007), in accordance with Rothengatter (1988) who mentioned that speed is consistent over time and locations, and in accordance with kinematic/dynamic analyses of steering and braking which show that speed is a crucial factor determining the critically of events (Allen et al., 2005b). Groeger (2000b) was one of the first to show that driving speed is consistent amongst learner drivers. Groeger made time measurements based on the number of video frames that elapsed between a start and endpoint on different road sections. Not only did this study show that learner drivers' number of speed limit violations increased with practice, it also showed that there were reliable correlations between drivers and road sections. Groeger concluded that "these results are, to the best of our knowledge, unique in showing a reliable relationship between chosen speeds across different time periods and road conditions, and suggest therefore, that some stable individual characteristic underlies the speeds at which drivers choose to drive" (p. 148). The present study supports and extends on the results of Groeger by using task completion time measurements of a diverse range of driving tasks (i.e., not only driving speed on straight road sections but also speed of shifting gears, using the indicator, etc.) and in diverse driving sessions (i.e., rural roads, urban areas, highways).

The importance of speed also applies to the assessment of surgeons' skills. In a recent study, we used principle component analysis (a technique highly similar to EFA) on surgeons' performance of various elementary psychomotor tasks (Chmarra et al., 2008). Interestingly, a variety of performance measures clustered onto the first principal component. We found the highest rotated loadings for the task completion time, and therefore we interpreted this component as *speed*. The extracted speed component had strong power to discriminate between experts, residents, and novice surgeons. More generally, this corresponds to previous studies into human skill acquisition that have shown that task completion time is an important characteristic (Crossman, 1959).

Participants with a higher speed-score made more violations but fewer errors (chapters 3 and 4). This is remarkable when considering that quicker participants performed more tasks than slower participants, and that quicker driving induces higher task demands. In other words, speed has a negative and a positive side: A swift task execution is indicative of vehicle control proficiency and good information-processing capacity. However, some drivers excessively prefer speed to accuracy and use their skills the wrong way. Literature indicates that when drivers use their skills to commit traffic violations, the results are likely to be negative for road safety (Hatakka et al., 2002). Others have also reported on the confounding relationship between violations and skills. Clarke et al. (2005) found that "specific groups of young drivers can even be considered as above average in driving skills, but simultaneously have a higher accident involvement due to their voluntary decisions to take risks" (p. 523). Reason et al. (1990) stated that "the finding that the subjects who report the most violations also tend to rate themselves as particularly skilful drivers suggests that these subjects believe that a good driver is someone who can 'bend the rules'" (p. 1330).

Allen et al. (2006) described the relationship between speed and performance in a large sample of teens that completed a training programme in a driving simulator. These authors found that students increased driving speed with increasing experience, as we found as well (chapters 2 and 3). However, mean speeds dropped when students had to pass a test without making collisions. Allen et al. (2006) attributed this phenomenon to a speed-accuracy trade-off: When the students had to perform accurately to pass the test, they "traded" speed for accuracy. These results are in accordance with the notion that car driving is a self-paced task in which drivers exert free control over task demands, and in accordance with our discussions about speed and accuracy in chapters 3 and 4. Salthouse (1979) was one of the first to investigate the extent to which individuals emphasize speed and accuracy in elementary psychomotor tasks. The speed accuracy distinction has been more formally introduced into the domain of car driving by Zhai et al. (2004), who extended Fitts' law of human movement into a steering law describing lane keeping performance. The results of Allen et al. may pinpoint limitations of driving tests: in case that tests emphasize accuracy they may not be a good representation of what drivers normally do.

2.5. Advancing simulation-based driver training

Chapter 5 provided recommendations to improve the effectiveness of simulation-based driver training. In addition, this chapter described how the results of chapters 2–4 can be used, and are already used, in Dutch driver training curricula. By summarizing previous data analyses, we recommended that students need didactically sound and clear-cut feedback and instructions during a driving lesson. A simple form of student-adaptive training by virtue of automatic regressive instruction was effective for learning to drive away. We recommended being cautious with augmenting the simulator with complex artificial intelligence. Moreover, an experiment showed that a lane keeping task can be effectively learned without feedback other than the task-intrinsic visual information. According to Groeger and Banks (2007), feedback during driving can be disruptive, which may be especially true for novices as they lack mental spare cognitive capacity as compared to experienced drivers. Lack of spare capacity was also used to explain the no-result of the learning of the mirror-checking task amongst inexperienced drivers in chapter 9. Chapter 5 concluded that it is crucial to determine in which situations the students actually need feedback and instructions. An alternative to providing feedback during task execution is to provide feedback after a training session has been completed or in-between the execution of critical and demanding driving tasks.

Chapter 5 proposed constructing a student-profile, with violations, errors, and speed as three primary factors. A basic form of a student-profile is already being used in Dutch Driving Simulators in the form of a strength-weakness report (chapter 5; De Winter et al., 2006b), as well as an online competition (Green Dino, 2008), both based on norm-referenced assessment. The purpose of a student-profile is to use it for getting a better understanding of the student. For example, human supervisors can draw on a student-profile to determine whether the student needs extra training or not, or whether a student needs additional tutoring to reduce tendencies for violating behaviour. Moreover, the (norm-referenced) scores can be used for competency-based training, requiring the students to achieve scores that are indicative of good and safe driving. This is already applied in Dutch Driving Simulators, where students repeat a lesson when their norm-referenced score is lower than a certain threshold (typically a 6 on a scale from 1 to 10; see also chapter 5), meaning that his or her performance is below a defined population percentile level. As explained, factor analysis can be used for calculating composite scores given a participant's data. However, as outlined in the discussion section of chapter 4, every single task still carries a certain level of uniqueness. This means that the factor scores should not be used as substitutes for individual task performance (see also chapter 5), and that training of individual tasks should not be discarded. Instead, the composite scores can provide an overall impression of the student with respect to the generic indicators speed, violations, and errors. An analogy can be made with student assessment in regular schools. Typically, a composite score such as a grade point average (GPA) or a (factor analysis based) intelligence test or personality test can be an excellent generic predictive-valid indicator for future job performance. Still, specific variables, such as students' performance on specific topics could be more predictive for specific types of job performance. The downside is that these specific variables are less generalisable and more dependent on the type of task and the type of assessment. Future research could also incorporate individual characteristics such as sensation seeking behaviour and personality into the student-profile. Research is needed to establish whether feedback and instructions that are tailored towards the individual strengths and weaknesses will improve training effectiveness. The suppression of high violation-scores seems relevant in the framework of road safety, as we recommend in section 3.2 as well.

2.6. Learning versus performance

The present thesis was largely devoted to scrutinizing how learners differ in simulation-based driving, leading to the extraction of speed-, violation-, and error-scores. Chapter 5 alluded to stable individual characteristics, such as personality, intelligence, and sensation seeking behaviour, which can also be included in a student-profile. However, a simulator should facilitate learning. Why did we focus on individual differences in performance, and not on individual differences in how people learn and acquire new skills?

First, in all the experiments and analyses, we found that individual differences in performance constitute a large share of the total variance. Individual differences were often much larger than the effects of feedback interventions (e.g., Figures 3 and 4 in chapter 7; see also Van Emmerik, 2004). Figure 2 of chapter 2 visually illustrated that differences between participants were larger than the effects of driving experience between sessions, and that these differences did not level out with increase of experience. Therefore, it seemed valuable to understand how and why people differ in their performance, rather than to focus on how people differ in their learning (i.e., a derivative of performance).

Second, we found that there is a conundrum in determining how much an individual has learned, because learning is related to the level of performance. For example, the learning curve of an expert driver will be virtually flat when a plateau is reached (i.e., a perfect score on a particular driving task), whereas a beginner may appear as a fast learner because he or she has a lot to learn. Consequently, it is difficult to make a clear-cut distinction between a person's learning pace and performance level. From our analyses of driving simulator data (not included in the present thesis), we did not identify a factor resembling individual learning characteristics. Moreover, we found that it is difficult to assess reliably the learning pace of a participant: Because learning is a derivative of performance, it has lower statistical reliability.

Third, there has been criticism regarding the concepts *learning styles* and *cognitive styles*. A literature study can easily show that applying learning/cognitive styles to individualize the training is highly popular, especially in the context of information technology. At least 70 different learning/cognitive styles have been put forward in the literature. A recent review of Menaker and Coleman (2007) aimed to distinguish between what has been empirically proven about learning/cognitive styles and what has popular appeal alone. Their conclusion was that it is likely that individuals differ in many ways, but there is a lack of empirical evidence linking learning/cognitive styles to learning outcomes.

We are not implying that the analysis of training systems should focus on performance instead of learning. To the contrary, analyses of learning (see chapter 5) are important to improve training effectiveness. The field of simulation-based training is in need for information about which feedback and instruction methods yield most satisfactory learning outcomes, and to what extent these results transfer from the simulator to the roads (see section 3.5). However, we recommend that student-adaptive training should guide towards reliable and valid individual traits, rather than towards traits that have inadequate psychometric properties.

2.7. Fidelity requirements

Chapter 6 reported on a literature study about the relationships between driving simulator fidelity and training effectiveness. The literature study provided a framework that outlined advantages, disadvantages, and unknowns that are associated with an intervention that increases simulator fidelity. Results showed that fidelity requirements are dependent on a compromise, in which negative effects of the intervention, such as costs, simulator sickness, distraction, and cue artefacts, have to be weighted against positive effects, such as improved validity, positive reactions, and better transfer of training. In addition, the study elucidated that understanding fidelity is an extremely complicated matter. Even qualifying and quantifying fidelity is difficult (Roza, 2005). Fidelity requirements also depend on the type of task and the simulator application (e.g., training vs. assessment, or experts vs. novices) (e.g., Kaptein et al., 1996).

The framework made clear that a more realistic simulator does not necessarily result in better training or assessment. This has been mentioned earlier in the literature (e.g., AGARD, 1980; Alessi, 2004; Kappé & Van Emmerik, 2005; Salas et al., 1998). Nonetheless, the assumption that *realism equates training effectiveness* strongly perseveres in the simulation community and is conveyed in many authorities' technical simulator requirements.

Using the framework of chapter 6, we recommended that improving the realism of force-feel characteristics of pedals is important. This relatively simple and inexpensive modification carries high functional relevance to the driving task.

2.8. Motion cueing systems

One of the main purported disadvantages of fixed-base simulators is that they lack tactile and vestibular cueing of vehicle accelerations, resulting in reduced validity of data (Greenberg et al., 2003). A possible solution to this deficiency is to implement a motion platform system. Referencing to the framework of chapter 6, one should carefully consider advantages, disadvantages, and unknowns before determining whether a motion platform is worthwhile.

The available literature shows that trainees generally prefer motion on to motion off. Furthermore, availability of motion cues generally improves operators' in-simulator performance, both in driving simulation (Mollenhauer, 2004), and flight simulation (Bürki-Cohen et al., 1998).[1] However, there is no scientific evidence that a simulator with motion platform helps to deliver more proficient pilots or drivers as compared to the same simulator without motion. The two meta-analyses that have been performed found statistically insignificant effects in flight simulation, one slightly negative (Hays et al., 1992), the other slightly positive (Vaden & Hall, 2005). Also, there is no solid evidence that the availability of a motion base prevents simulator sickness (McCauley, 2006; Mollenhauer, 2004). Finally, it is important to note that a traditional Stewart motion platform does not provide physically realistic accelerations. Studies have shown that there is a large discrepancy, both in magnitude and pattern, between in-cabin accelerations and vehicle model accelerations, caused by travel and bandwidth limitations of the motion system (Colombet et al., 2008; Tomaske et al., 2001). More expensive solutions, such as linear drives, are required to provide physically realistic motion cueing during a lane change, for example (Nordmark et al., 2004). To summarize, current motion platforms improve user acceptance, provide a perceptual illusion of motion, and elicit more realistic in-simulator performance by providing task-relevant information (which does not necessarily has to be realistic in terms of forces and accelerations).

Simple low-cost solutions such as vibration systems or small amplitude cueing may be an adequate resolution to the aforementioned compromise of advantages and disadvantages, while being able to satisfy the functions of more complex systems. The potential merit of low-cost solutions becomes apparent when reflecting on the costs of motion. As discussed by Vaden and Hall (2005): "the question is not just whether there is an advantage to having motion but how valuable any existing advantage may be for pilot training. One should consider some of the costs .... Limited availability (largely due to ownership costs) for certified simulators also means scheduling issues, travel costs, and time away from the job for many trainees .... Motion platforms require more physical space, more computing power, greater environmental control, more manpower for support, and result in higher maintenance

[1]) In the present reflection regarding motion cueing systems, reference is frequently made to flight simulation. This is because more research and expertise exists in this domain as compared to driving simulation. Flying and driving pose very different demands in terms of accelerations and functional motion cues. Note that requirements also differ between military vehicles and regular car driving. For instance, very little outside visual information is available in tanks, and mine resistant vehicles are less stable than normal cars because of their high centre of gravity; therefore the drivers of these military vehicles have to rely more on nonvisual cues. Nonetheless, the ideas and conclusions in this thesis may be generalizable to flying and military vehicles as well.

costs. Perhaps new motion chair devices will provide some or even all of the training benefits that full-motion platforms currently offer at a fraction of the price" (p. 389).

Chapters 7 and 8 experimentally tested the following seven low-cost systems: a motion seat, a seatbelt tensioning system, a stiff brake pedal, a vibrating steering wheel, screeching tyre sound, a vibrating seat, and a pressure seat. These systems aimed to provide the participant with additional information about vehicle accelerations. In later studies, not included in this thesis, we tested four other systems: a brake pedal incorporating a progressive spring and hysteresis, an infinitely stiff brake pedal that used a force sensor, auditory beeps that provided information on vehicle accelerations, and a vibrating seat for simulating road noise (De Groot et al., 2008). Considering that participants in simulators adopt accelerations that are uncomfortably high in reality, each system was hypothesized to lead to reduced speeds and accelerations of the virtual vehicle.

Figure 2 shows the consistency of the stopping distance for the different systems tested in our fixed-base driving simulator, as well as two systems described in the literature. It can be seen that the provision of additional cues improved stopping consistency in all cases (although not always statistically significant). Similarly, all the motion systems in chapters 7 and 8 resulted in reduced maximum decelerations and reduced cornering speeds, although the effects were statistically insignificant for the latter. Siegler et al. (2001) discussed the phenomenon of in-simulator performance improvement using the principle of sensorimotor integration: "motion cues – when present even with limited amplitude – may be integrated in the driver control loop as additional inputs, providing they are relevant to the driving task considered". In other words, drivers use the available information to improve their performance.

Simple modifications, such as the altered brake pedal stiffness, had notable effects on braking performance. In accordance with the framework of chapter 6, such simple task-relevant cues have to be optimized first: It does not seem sensible to spend considerable resources to a motion platform when a plain element such as the brake pedal has not been validated yet. In addition, results showed that the motion seat and the tensioning seatbelt had beneficial effects on the subjective realism of the simulator. It is appealing that a tensioning seatbelt can contribute to the feeling of being present in a real vehicle, while seatbelt tensioning forces do not exist in a real car. This suggests that humans can easily be provided with an illusion of motion. The pressure seat, on the other hand, was not successful in eliciting positive reactions. There were indications that it lacked sufficient stimulus-response compatibility so that participants were not able to use the cues to their benefit.

To summarize, the experiments showed that low-cost motion cueing systems can be successful in improving in-simulator performance and in increasing subjective realism. This is an important step in providing justification for low-cost motion cueing systems as substitutes for more complex systems.

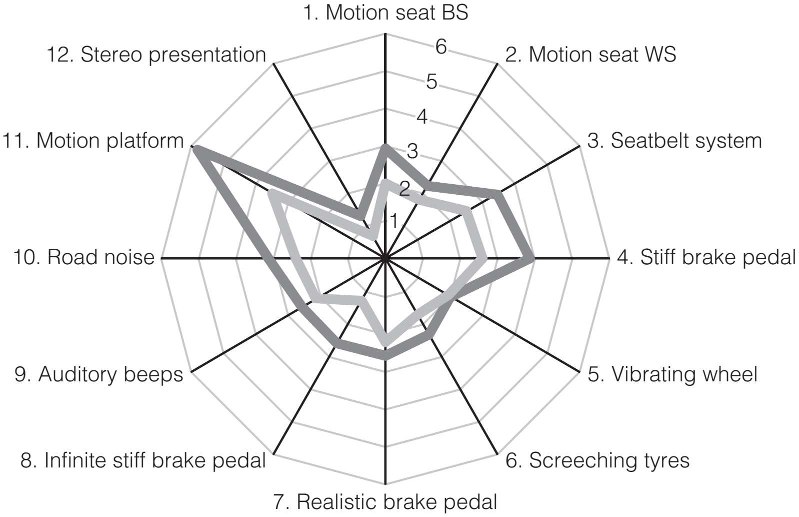


Figure 2. Mean of participants' standard deviation of distance to the target when standing still (i.e., stopping consistency, SD DTLfin [m]) for 12 driving simulator experiments in which participants had to stop at a target. The experimental condition is indicated in light grey; the control condition is indicated in dark grey. The following experiments are shown: *1) Motion seat between-subjects (chap. 7), 2) Motion seat within-subjects (chap. 7), 3) Seatbelt tensioning system (chap. 8), 4) Stiff brake pedal (chap. 8), 5) Vibrating steering wheel (chap. 8), 6) Screeching tyre sound (chap. 8), 7) Progressive brake pedal with hysteresis (De Groot et al., 2008), 8) Infinite stiff brake pedal (De Groot et al., 2008) 9) Auditory beeps depending on vehicle acceleration (De Groot et al., 2008), 10) Speed-dependent seat vibrations or "road noise" (De Groot et al., 2008), 11) Limited amplitude motion platform (Siegler et al., 2001), 12) Stereo presentation (Kim et al., 2005).* Note that experiments 1 and 2 have participant overlap (see chap. 7). The treatments of experiments 7 and 8 were actually the same between-subjects experiment; the SD DTLfin of the treatment of experiment 4 was used as a substitute for the control group in experiments 7 and 8, because it was the simulator in the same configuration. Experiment 11 represents an aggregate SD DTLfin of all participants. Note that a driving simulator study of Pinto et al. (2004) showed that visual tilt cues as well as motion cueing improved stopping accuracy to a target. Unfortunately, this study included insufficient quantitative information to be included in the present figure.

The motion cueing experiments contributed to fundamental thinking about simulator fidelity in relation to training effectiveness. It was shown that there are many ways to improve drivers' in-simulator performance so that it better matches real world driving behaviour. The human brain uses efficient – often statistically optimal – strategies to integrate and interpret information (Cutting, 1997; Ernst & Bülthoff, 2004). In essence, trying to resemble performance in a simulator to performance in the operational environment (i.e., having accurate behavioural fidelity) is an underdetermined problem. This could mean that performance in the simulator (speeds, accelerations, etc.) can accurately match performance in the operational environment by using “unrealistic” cues. Similarly, it could be possible that two simulators show identical driver performance, but both elicit dissimilar driver’s cue weighting strategies.

Further research is needed to establish whether (a combination of) motion cueing systems contributes positively to training effectiveness. Humans seem well able to reschedule the weights of different information sources. For example, research has shown that humans are capable of adapting to various system dynamics (McRuer & Jex, 1967), to cars with different steering wheel gains (Evans, 2004), or to driving with a time delay (Cunningham et al., 2001). This could imply that effective transfer takes place even if the simulator cues are only partially realistic. A danger, however, is that insufficient transfer occurs because the human has only learned to master the vehicle in the simulator. That is, he or she cannot apply the results out of the context of virtual reality. There exist little empirical studies investigating the degree of transfer from driving simulation to the operational environment. One of the few examples is Uhr et al. (2003), who showed that both positive and negative transfer can take place from the simulator to the roads. Recommendations for simulator validation are provided in section 3.5.

2.9. Limitations

Several limitations apply to the experiments in chapters 5, 7, 8, and 9. The experiments were performed with participants from a technical university. This provided us with a suitable pool of youngsters, many of them without a driving licence. However, the university population has specific sociodemographic characteristics: The participants were predominantly men with a profound interest in technology. There were no indications that the participants in the experiments showed social desirability in the questionnaire responses. To the contrary, participants were critical towards the realism of the simulator with mean ratings between about 4 and 6 on a scale from 1 to 10 (e.g., chapters 7–9). Moreover, not all systems increased user ratings (e.g., pressure seat in chapter 8 and mirror-checking feedback in chapter 9). An interesting result was that participants’ ratings of didactic aspects of the simulator were higher: between 6.5 and 8.5 on a scale from 1 to 10 (chapter 9 for university students; Van der Snee, 2005 for interviews at Dutch driving schools). An alternative to university students is to acquire participants who respond to an advertisement or to recruit them from a database of volunteers. Such a method was used in Houtenbos (2008) and De Winter et al. (2006a). The risk that looms for this approach is that the sample consists of participants who are very interested in scientific research and who may start thinking about the design of the experiment. To summarize, behaviour during the experiments in chapters 5, 7, 8, and 9 may not be fully representative of behaviour in driving schools where students are paying for their lessons in order to obtain their licence.

The experiments of chapters 7 and 8 were of different designs. Simulator configurations varied as well. Some experiments included licensed drivers and others included unlicensed drivers; some used a between-subjects design and others a within-subjects design. The advantage of a between-subjects design is that conditions are independent and that there are no carryover effects. The advantage of a within-subjects design is increased statistical power, but a drawback is that students can more easily become aware of the aim of the experiment, which may also lead to inflated effect sizes. Consequently, the results of each independent experiment were valid, but the results of different experiments should not be compared on an absolute scale. Nonetheless, in a more detailed analysis provided by De Groot et al. (2008), it was found that the results of chapters 7 and 8 are reasonably invariant for subsamples, such as for experienced versus inexperienced participants. The result that men adopt higher speeds and more accurate lane keeping performance has been found in driving schools as well as in experiments in our laboratory, thereby providing support for the generalizability of this result.

The experiments were complemented by studies that had recorded students' performance during simulation-based training in driving schools in the Netherlands (chapters 3 and 4). This provided important information about how students behave in a naturalistic setting and provided excellent statistical power due to large sample sizes. An important weakness, however, is that we lacked experimental control over who drove in the simulator. For example, chapter 4 found that students who completed more simulator lessons obtained their driving licence at an earlier date. Because of the lack of a control group, it is impossible to conclude that simulator training *causes* earlier licensure or whether it was a spurious correlation arising from volunteer bias.

The present thesis quantified students' *higher order behaviours* (i.e., violations and speed choice) as well as *lower level skills* (i.e., errors, vehicle control). However, it takes more to improve training effectiveness. A more pragmatic approach is needed to investigate how to train and assess specific tasks and procedures. For this, one should make use of task analyses (e.g., McKnight & Adams, 1970) and learning objectives (Vlakveld, 2000). Further recommendations on simulator didactics are provided in section 3.1.

In this project, we assessed the predictive validity of driving simulator scores on driving test results. As indicated in chapter 4, it would have been relevant as well to assess predictive validity with respect to a driver's crash involvement. Unfortunately, during the course of this project, we could not acquire police-recorded crash data of the students. It may have been feasible to send everyone (of whom the post address was known) a letter asking for the number of crashes since licensure. However, due to limited response rates, volunteer bias in the type of respondents (cf.

Hazevoet & Vissers, 2004 for a questionnaire study amongst driving exam candidates), and biases in the accuracy of subjective crash record data (e.g., Ranney, 1994; Groeger, 2000), we reasoned that these crash data would be of insufficient validity and insufficient statistical power anyway.

Another limitation of the present research, with the exception of chapter 4, is that it concentrated on in-simulator performance. As discussed in chapter 3, each simulator provides an imperfect representation of real driving. Safety has a different meaning and the consequences of committing violations are different compared to reality. It is vital to have information about the extent to which behaviour in a simulator *predicts* future crash involvement. In addition, it is relevant to investigate the extent to which simulator training *affects* later on-road driving. Better insight in these matters can aid in specifying fidelity requirements and can support software developments. Further recommendations are provided in section 3.5.

A limitation of the present research about simulator fidelity is that – with the exception of stereo presentations in chapter 6 – it was restricted to motion cueing. The importance of visual and auditory information should not be neglected. Studies have shown that visual information alone can yield an illusion of motion (Riecke et al., 2005). A personal experience on a ship simulator showed that a large field of view projection yields a very compelling illusion of tilt motion in the absence of any physical motion. Auditory feedback can influence speed perception as well as accuracy of vehicle control (Evans, 1970; Pinto et al., 2004). Future research should evaluate the extent to which a driver uses these different modalities, and to what extent they contribute to training effectiveness.

3. Future research directions

The chapters in this thesis provided in-depth recommendations. Here, the focus is on what we consider the most important suggestions for future research in the field of simulation-based driver training.

3.1. Improving feedback and instructions

It is important to investigate further which types of feedback and instruction generate the best possible training effectiveness. It has been mentioned in chapters 5 and 9 that potentially large gains can be made by improving simulator didactics, which can likely be achieved with simple rather than complex interventions. Important aspects to be considered are the implementation of the *Five Principles of Instruction* (De Groot et al., 2007; Merrill, 2002), providing augmented feedback (De Groot et al., 2006; Van Emmerik, 2004), considering in which situations students actually need feedback or instructions (see also chapter 5), and conducting extensive software tests by means of video observation.

A tentative suggestion is to conduct large-scale experiments in driving schools. Here, the simulator can randomly assign the students to a treatment or a control condition, without informing the student or the human supervisors. In the end of the simulator curriculum, a simulator-driving test can be used that is equal to all students. Such studies may be an efficient way to test the effects of various modes of feedback and instructions in an ecologically valid setting with excellent statistical power, yielding important results for the scientific community.

In this respect, it is relevant to compare different training philosophies. For example, should training include many critical hazards (such as triggered vehicle and pedestrian encounters) for learning behavioural compensation and defensive driving, or would it be more safety-effective to not include such triggered critical hazards and to focus on common normative driving behaviours instead? Similarly, it is important to investigate the effectiveness of part-task training versus whole-task training, and the effectiveness of training of lower-order vehicle control versus focusing on higher-order skills and attitudes.

3.2. Remedying violations

This thesis showed that a simulator is able to provide individual violation-scores before the students have actually driven a real car. The importance of early identification of deviant behaviours was substantiated by a recent study from the United Kingdom, which concluded that "drivers tended to enter the driving population with fairly fixed ideas about themselves – both in absolute terms and in relation to others. This indicates that interventions to influence attitudes need to be in place very early in a learner driver's training, or even prior to practical training" (Wells et al., 2008, p. 12).

Chapters 3, 4, and 5 discussed possibilities for remedying violations. An important countermeasure could be deflation of driver confidence (Senserrick, 2001) and multimedia campaigns to change attitudes and norms. However, some have argued that an individual's crash risk is governed by strong developmental and biological factors (Arnett, 2002; Evans, 2006). Indeed, looking at the consistency and robustness of the speed with which people drive through the virtual environment (e.g., chapters 5 and 7), it may be argued that remedying violations will require engineering and enforcement solutions as well, such as speed limiters and stronger police enforcement.

3.3. Simulator sickness

Participants in our experiments had little problems with simulator sickness symptoms. This can be explained by the fact that the participants were mostly inexperienced drivers and that our experiments consisted of short-lasting simple driving tasks (e.g., Mollenhauer, 2004). Simulator sickness can be a more serious problem undermining the validity and credibility of the system when involving experienced drivers in demanding driving tasks.

Alleviation of simulator sickness is unlikely unless a valid underlying model can explain its causes. Scientific literature has put forward numerous theories of simulator sickness, with the sensory conflict (SC) theory and the postural instability (PI) theory as the prominent ones (Johnson, 2005; Mollenhauer, 2004; Zaychik & Cardullo, 2003). However, these existing theories are not without weaknesses. The SC has been criticized for having little predictive power (Johnson, 2005). Furthermore, SC and PI make different predictions regarding the relative contributions of the confounding variables age and experience (Johnson, 2005). Understanding the symptoms of simulator sickness is largely based on EFA (Kennedy et al., 1993). Future research into simulator sickness should develop a quantitative model. Having a valid model can better help to identify risk factors and to produce methods that alleviate simulator sickness, which is of great importance to the simulation community.

3.4. Driver behaviour modelling and statistics

Having a good understanding of (young) drivers' behaviour is important for road safety. The present thesis showed that simulators can contribute to such understanding. In addition, this thesis stressed the importance of data collection in simulation-based training. However, the present ideas for driver modelling do not necessarily restrict itself to simulation. Interesting possibilities arise when considering the increasing availability of driver surveillance systems on the roads. Moreover, analyses of coupled databases, such as those of the driving test organisations, simulator performance records, and accident records, will likely yield vital results for driver behaviour modelling.

One may ask why we proposed EFA for modelling the driver and not another statistical technique. The reason to prefer EFA is that it allows a parsimonious representation of the data by virtue of latent factors. When using techniques such as regression analysis or linear discriminant analysis (LDA) directly on the manifest variables, there is a higher risk of overfitting and multicollinearity, resulting in a model that does not adequately capture the underlying phenomena. Therefore, and particularly when there are many intercorrelated variables, it is recommended to conduct LDA, regression analysis, or correlation analysis, only after performing EFA.

A number of statistical methods carry close similarity to EFA. Although researchers often claim that EFA and principal component analysis (PCA) are different methods with different aims, in practice they yield highly similar results (Velicer & Jackson, 1990). Another appealing technique, closely related to EFA and PCA, is independent component analysis (ICA). A powerful feature of ICA is that it can be used for rotation towards independence rather than towards a simple structure, as is normally done on the loadings matrix acquired with EFA or PCA (Jennrich & Trendafilov, 2005). Confirmatory factor analysis and structural equation modelling – techniques related to EFA – could be valuable as well for future driver behaviour modelling research.

3.5. Transfer of training and predictive validity

Technological advancement has been of crucial importance in the development of any vehicle simulator. However, the driving simulation community should shift its focus from introducing technological novelties and direct more attention towards data collection and validation. It is recommended to regard a simulator as a tool for training and assessment, rather than a replication of the operational environment, as the term *simulator* implies. Successful validation of driving simulators against more traditional modes of training could evoke a more widespread acceptance of the former. Validation and transfer studies are important but relatively scarce in the present field of simulation-based driver training.

Most professional on-road driver training methods do not positively contribute to road safety as compared to informal training (chapter 1). However, there has been no research investigating the effectiveness of simulation-based pre-licence driver training. As previously outlined by Vlakveld (2006b), it is important to determine whether training in a simulator improves road safety. Answering this question requires a large and well-designed experiment that randomly assigns participants between groups. For sufficient statistical power, probably at least 500 participants will be needed (cf. chapter 4). However, 10,000 or more is a better number to aim at, following the design of the DeKalb study (Lund et al., 1986). It is necessary to have full experimental control; cross-sectional studies do not suffice because of volunteer bias. In addition, support from driving test organizations as well as organizations that have access to drivers' accident records is recommended.

Chapter 4 investigated the extent to which driving simulator performance can predict driving test results. It is also relevant to investigate whether a driving simulator can predict future crash involvement. Specifically, future research should evaluate whether a simulation-based test is better able to discriminate between safe and unsafe drivers than an on-road test with a human examiner, or than a battery of neuropsychological tests. The hypothesis that a simulator test is better able to identify crash-prone drivers than a human examiner is certainly justified when considering the dissatisfactory reliability of the driving test (Baughan et al., 2005).

The use of driving simulators for testing is increasingly recognized in the clinical domain and the assessment of older drivers (e.g., George, 2003; Lew et al., 2005). An experiment by Lew et al. (2005) concluded that "simulator-based assessment of patients with brain injuries can provide ecologically valid measures that, in some cases, may be more sensitive than a traditional road test as predictors of long-term driving performance in the community" (p. 177). Recently, simulators have been allowed for recurrent training and testing of truck and bus drivers (CCV, 2008). It can certainly be expected that simulators will get a more prominent role in formal training and testing of learner drivers, not unlike current training and assessment of airline pilots, which for a large part takes place in simulators. The results of the present thesis showed that simulation-based assessments do not have to constrain to skills and procedures; they can also be used to assess driving style aspects. It is therefore recommended that a simulation-based test should incorporate a diverse range of tasks, prolonged periods of self-paced driving, and scenario's that can bring about violating behaviour such as speeding. Standardization is important in order to facilitate valid iterative decision-making with respect to tasks that should be included, as well as for comparing different types of simulator configurations. A breakthrough is envisaged when future research is able to show that a simulation-based test yields data that are more valid and more reliable than traditional tests with a human examiner.

It is relevant as well to compare different types of simulators (i.e., cross-platform comparisons) with respect to their training effectiveness and their predictive validity. Especially the comparison between a simulator with and without a motion platform seems important (see Bürki-Cohen et al., 1998, and Vaden & Hall, 2005, for recommendations in flight simulation). An interesting (and perhaps the only) study into the predictive validity of motion cueing was performed by Koonce (1979). This study evaluated three conditions in a flight simulator: (a) no motion, (b) simple sustained linear motion, and (c) a more sophisticated washout motion, using 90 licensed multi-engine pilots. Results showed that no motion yielded the poorest and washout motion the best in-simulator performance. There were no significant differences in the real aircraft, although the no motion group tended to perform best in the contact manoeuvres. The correlations (i.e., predictive validity) between pilot performance in the simulator and pilot performance in the aircraft were 0.76 for no motion, 0.91 for sustained motion, and 0.65 for washout motion. Koonce (1979) concluded that for predictive validity, it is important to consider which motion system provides the best stability in performance, rather than which system provides the best fidelity of motion cues or which system provides the best in-simulator performance. The washout motion system caused variation within and between individuals, thereby reducing predictive validity. As already discussed in chapter 4, a high-fidelity simulator does not necessarily yield better data than a low-fidelity one. Instead, simplicity and standardization are important features in order to obtain reliable measures.

To summarize, it is recommended to put more research effort into investigating to what extent the training in a simulator *affects* and *predicts* future driving on the road. Here lies considerable potential to the benefit of safety on the road and society as a whole.

APPENDIX A

Gender differences in driving licence theory test scores in the Netherlands

Abstract

Gender differences were investigated in a sample of persons (*N* = 34,775) who completed the driving licence theory test in the Netherlands. Contrary to recent findings from Sweden, no gender differences were found. The present study signifies the importance of standardization in driver testing.

De Winter, J.C.F., & Wieringa, P.A. (2008). Gender differences in driver's license theory test scores in the Netherlands. *Journal of Safety Research*, *39*, 413–416. (adapted with minor textual changes)

1. Introduction

It is a well-known phenomenon that young male drivers are more involved in severe car crashes as compared to their female counterparts. To obtain better insight into gender differences amongst young drivers, Wiberg (2006) investigated the results of 11,862 persons who completed the Swedish driving licence theory test in the period January 9, 2004 to February 9, 2004. She found that men had a mean score of 52.70 out of a maximum score of 65 as compared to 54.19 for women. The author considered that the inferior performance of men in the theory test could be a factor that explains why male drivers are overrepresented in road traffic crashes.

Inspired by the noteworthy findings of Wiberg (2006), we investigated whether gender differences in theory test results exist in the Netherlands as well. Additionally, we looked at the mediating role of age. For this purpose, we analysed the scores of 34,775 persons who completed the Dutch driving licence theory test in the period September 8, 2003 to November 12, 2003.

2. Theory test procedures in the Netherlands

All theory tests were taken at one of the 32 theory-testing locations of the Dutch Driving Test Organization (CBR). For the period under investigation, the Dutch theory test comprised 50 computerized questions: yes/no questions, multiple choice questions, and open questions (in which the test taker should type in a number). The questions were asked orally based on images of traffic situations shown on large TV screens. Candidates were also able to read the questions on these TV screens. Candidates had to give answers by means of buttons on the desk. To pass the test, participants needed a minimum of 45 correct answers.

After successfully completing the theory test, a theory certificate is acquired which is valid for a period of 1 year. The learner driver must have a valid theory certificate before being allowed to participate in the practical driving test. In the Netherlands, people are allowed to have on-road driving lessons and undertake the theory test only after turning 18 years old.

3. Sample under investigation

Results of 34,775 persons who passed the theory test in the period September 8, 2003 – November 12, 2003 were obtained from the CBR. Test results of failed tests (score 44 or lower) were not available. Note that, according to the CBR annual report (CBR, 2003), 202,680 of 410,779 participants passed the theory test in that year.

Table 1. Gender differences in test scores and age when passing the theory test in the Netherlands

	Men (n = 16,627)			Women (n = 18,148)		
	Mean	SD	95% CI	Mean	SD	95% CI
Test score [45–50]	46.69	1.394	46.67–46.71	46.70	1.397	46.68–46.72
Age [years]	21.12	5.641	21.03–21.21	22.69	7.373	22.58–22.79

Note. Standard deviations (SD) and 95% confidence intervals (95% CI) were calculated assuming normal distribution. It should be noted, however, that the distribution of age was highly skewed.

4. Results

Table 1 shows the mean score of men and women. An independent *t* test indicated that there was no statistically significant gender difference (p = 0.7). Figure 1 shows the distribution of test scores of men and women. Again, no significant gender difference was identified.

Table 1 also shows that men had a lower mean age than women when passing the test (p < 0.001 using a *t* test). Note that this difference was larger than the age difference amongst theory test-takers in Sweden (Wiberg, 2006), where mean ages of 19.64 and 19.85 years were found for men and women, respectably (overall SD = 1.54 years). Figure 2 illustrates the relationship between age and gender. It

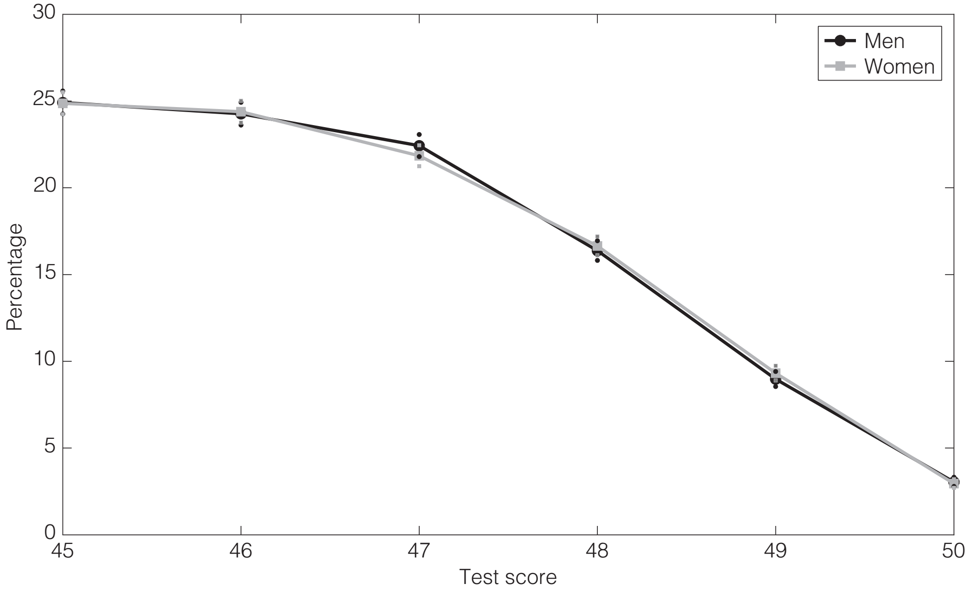


Figure 1. Distribution of theory test scores for men and women. 95% confidence intervals were calculated for each score separately assuming binomial distribution.

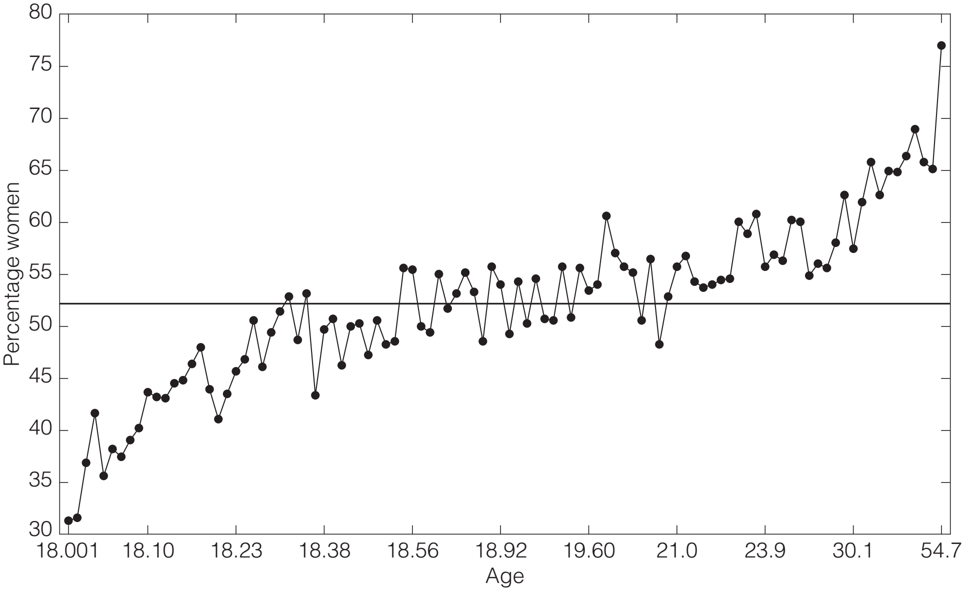


Figure 2. Proportion of women versus age when passing the theory test. The graph was created by sorting the persons according to their age and creating 100 groups of 347 or 348 persons each. Each point in the graph represents the percentage of women in the group versus the mean age of the group. The horizontal line depicts the proportion of women in the total sample (52.2%).

can be seen not only that the distribution of age is highly skewed, but also that those who passed the driving test at very early age were mostly men. Conversely, persons who were involved in driver education at an older age were predominantly women.

Figure 3 depicts the interaction between age and test score. Test score had a small significant negative correlation with age, both amongst men (Spearman correlation -0.0514, $p < 0.001$) and amongst women (Spearman correlation -0.0564, $p < 0.001$). We also calculated the mean score of men and the mean score of women for each of the 100 groups shown in Figure 3. A paired *t* test revealed no statistically significant difference ($p = 0.11$), indicating that no gender differences were identified after correcting for age.

5. Discussion

In the Netherlands – contrary to the findings from Sweden – no gender differences in theory test scores existed amongst those who passed the test. This discrepancy could imply that gender differences in theoretical knowledge truly differed between countries at the time of taking the test, or that the theory test of one or both countries was somehow gender biased, or that the tests were measuring different aspects of

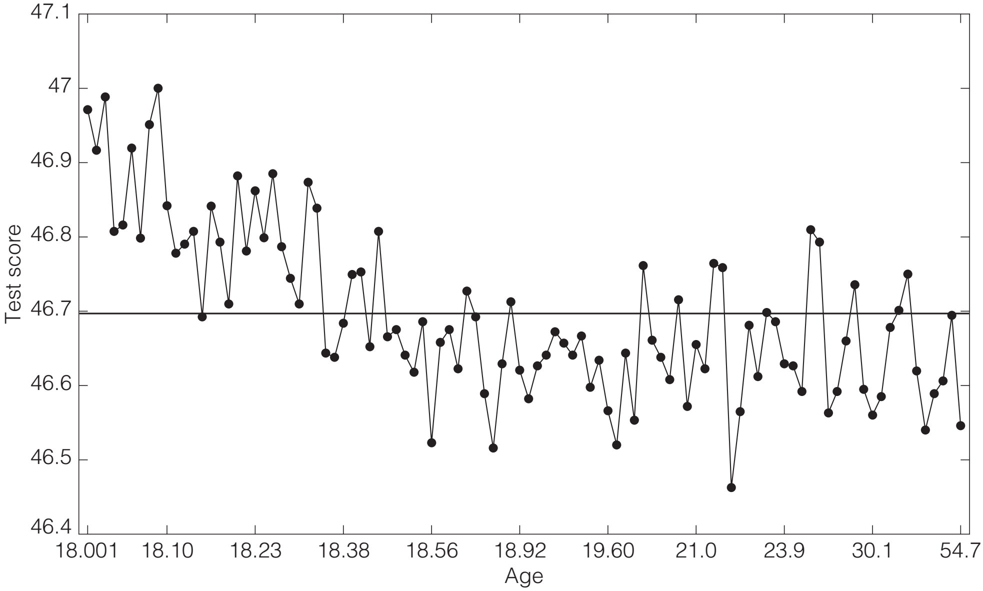


Figure 3. Test score versus age when passing the theory test. The graph was created by sorting the persons according to their age and creating 100 groups of 347 or 348 persons each. Each point in the graph represents the mean test score of the group versus the mean age of the group. The horizontal line depicts the mean test score of the total sample (46.70).

theoretical knowledge. Therefore, no definite conclusions can be drawn of the supposed relationship between theory test scores and crash rate. Regardless of the explanation, the present results signify the need for improved standardization in driver training and testing procedures, so that valid comparisons become possible and valid conclusions can be drawn with respect to gender differences in traffic safety (see Baughan et al., 2005, for further consideration).

The present study found no gender differences in theory test results when passing the test. Other studies have found no gender differences in the practical driving test (Nyberg et al., 2007; Wiberg, 2006, both in Sweden). Conversely, others have identified considerably higher pass rates amongst men in the practical driving test (e.g., Crinson & Grayson, 2005, in the UK). Also, wide gender differences have been found in driving simulators (chapter 3, in the Netherlands). The need for further research into gender differences in car driving was underscored by the World Health Organization (2002), which also stressed the need for standardization of data collection procedures.

A final remark is made with respect to effects of volunteer bias. Figures 2 and 3 suggested that volunteer bias effects can play an important role in driver statistics. It is unlikely that the decrease of test scores within a couple of months time reflects a true effect of age. It is much more likely that (predominantly male) drivers who obtained their licence quickly after their 18th birthday had a very high motivation toward driving. According to Hatakka et al. (2002), a highly car-oriented lifestyle could be particularly problematic for road safety. The present study illustrated that because of volunteer bias effects, researchers should be cautious in interpreting driver-related statistics, for instance, in quantifying the effects of age on crash risk.

APPENDIX B

Exploratory factor analysis with small sample sizes

Abstract

Exploratory factor analysis (EFA) is generally regarded as a technique for large sample sizes (N), with $N = 50$ as a reasonable absolute minimum. This study offers a comprehensive overview of the conditions in which EFA can yield good quality results for N below 50. Simulations were carried out to estimate the minimum required N for different levels of loadings (λ), number of factors (f), and number of variables (p), and to examine the extent to which a small N solution can sustain the presence of small distortions such as interfactor correlations, model error, secondary loadings, unequal loadings, and unequal p/f. Factor recovery was assessed in terms of pattern congruence coefficients, factor score correlations, Heywood cases, and the gap size between eigenvalues. A subsampling study was also conducted on a psychological dataset of individuals who filled in a Big Five Inventory via the Internet. Results showed that when data are well-conditioned (i.e., high λ, low f, high p), EFA can yield reliable results for N well below 50, even in the presence of small distortions. Such conditions may be uncommon but should certainly not be ruled out in behavioural research data.

De Winter, J.C.F., Dodou, D., & Wieringa, P.A. (in press-b). Exploratory factor analysis with small sample sizes. *Multivariate Behavioral Research*. (adapted with minor textual changes)

1. Introduction

Exploratory factor analysis (EFA) is one of the most widely used statistical methods in psychological research (Fabrigar et al., 1999), prompted by the need to go beyond the individual items of tests and questionnaires to reveal the latent structure that underlies them. Factor analyses are generally performed with large sample sizes. A study of the literature easily shows that applying EFA to small sample sizes is treated with caution. Researchers are discouraged from using EFA when their sample size (*N*) is too small to conform to the norms presented in the state of the art in factor analysis. Many early recommendations focused on the importance of absolute sample size. Guilford (1954) recommended a minimum sample size of 200 for consistent factor recovery. Comrey (1973) suggested a range of minimum sample sizes, from 50 (very poor) to 1,000 (excellent), and advised researchers to obtain sample sizes larger than 500. Gorsuch (1974) characterized sample sizes above 200 as large and below 50 as small. Cattell (1978) proposed that 500 would be a good sample size to aim at, commenting that in the context of most problems, however, 250 or 200 could be acceptable. Other researchers focused on the number of cases per variable (*N*/*p*) and recommendations range from 3:1 – 6:1 (Cattell, 1978) to 20:1 (Hair et al., 1979), with the latter advising researchers to obtain the highest cases-per-variable ratio possible in order to minimize the chance of overfitting the data.

Later studies showed that those propositions were inconsistent (Arrindell & Van der Ende, 1985) and recommendations on absolute *N* and the *N*/*p* ratio have gradually been abandoned as misconceived (Jackson, 2001; MacCallum et al., 1999). Meanwhile, a number of studies have pointed out that not only sample size but also high communalities (Acito & Anderson, 1980; Pennell, 1968) as well as a large number of variables per factor (*p*/*f*) (Browne, 1968; Tucker et al., 1969) contribute positively to factor recovery. Recently, a steeply increasing number of simulation studies has investigated the determinants of reliable factor recovery and shown that minimum sample size is a function of several parameters. There are no absolute thresholds: Minimum sample size varies depending on the level of communalities, loadings, number of variables per factor, and the number of factors (Gagné & Hancock, 2006; MacCallum et al., 1999, 2001; Marsh et al., 1998; Velicer & Fava, 1998). A considerable part of the literature on sample size recommendations has been reviewed by Velicer and Fava (1998) and MacCallum et al. (1999).

MacCallum et al. (1999) developed a theoretical framework for the effects of sample size on factor recovery and provided a basis for the contention that there are no absolute thresholds for a minimum sample size. This framework is based on earlier theoretical analyses presented by MacCallum and Tucker (1991), subsequently extended by MacCallum et al. (2001). The framework indicates that factor recovery improves as: (a) sample size increases, (b) communalities increase, and (c) *p*/*f* increases; the effect of *p*/*f* decreases as communalities increase, and it may also interact with the sample size. Although the simulations in MacCallum et al. (1999, 2001) applied a minimum *N* of 60, their theoretical framework should be applicable to smaller sample sizes as well. However, it remains undefined how small a sample size can be and still yield acceptable solutions.

Only a very limited number of studies on the role of sample size in factor analysis have investigated real or simulated samples sized smaller than 50, probably because this is considered a reasonable absolute minimum threshold (Velicer & Fava, 1998). A few earlier studies recognized that sample sizes of 30 (Geweke & Singleton, 1980, having tested sample sizes as small as 10) or 25 (Bearden et al., 1982) can be adequate but, as Anderson and Gerbing (1984) noted, the latter study was limited and its findings should not be generalized. In a Monte Carlo study on confirmatory factor analysis (CFA) with sample sizes ranging from 25 to 400, Boomsma (1982) characterized factor analysing sample sizes smaller than 100 as "dangerous" and recommended using sample sizes larger than 200 for safe conclusions. A subsampling study of Costello and Osborne (2005) indicated that for a sample size as small as 26, only 10% of the samples recovered the correct factor structure, whereas 30% of the analyses failed to converge and 15% had Heywood cases. A study by Marsh and Hau (1999) specifically devoted to small sample sizes in CFA used a minimum of 50 and warned that reducing the sample size from 100 to 50 can dramatically increase the number of improper solutions. Sapnas and Zeller (2002) determined adequate sample sizes for principal component analysis and suggested that a sample size between 50 and 100 was adequate to evaluate psychometric properties of measures of social constructs. This study, however, has been criticized for methodological errors and for failing to explain under which conditions a small sample EFA may be feasible (Knapp & Sawilowsky, 2005). In a more recent work, Zeller (2006) concluded that a sample size between 10 and 50 was sufficient for two dimensions and 20 variables. A simulation study by Preacher and MacCallum (2002) on applying EFA in behaviour genetics clearly showed that for communalities between 0.8 and 0.9 and 2 factors EFA can yield reliable solutions even for sample sizes as small as 10. Another recent Monte Carlo study by Mundfrom et al. (2005) also showed that if communalities are high and the number of factors is small factor analysis can be reliable for sample sizes well below 50. Finally, Gagné and Hancock (2006) found that a sample size of 25 yielded no incidences of nonconvergent or Heywood cases when loadings were as high as 0.8 and $p/f = 12$. The majority of these studies, however, did not investigate factor recovery when deviating from a simple structure; a situation most likely to be encountered in real data. An exception was the study by Preacher and MacCallum (2002), but it included model error as the sole distortion.

This paper aims to offer a comprehensive overview of the conditions in which EFA can yield good quality results for small sample sizes. A number of simulations were carried out to examine how the level of loadings and communalities, the number of factors, and the number of variables influence factor recovery and whether a small sample solution can sustain the presence of distortions such as interfactor correlation, model error, secondary loadings, unequal loadings, and unequal p/f. Next, we conducted a subsampling study on a psychological dataset of individuals who filled in the 44-item Big Five Inventory. The dataset was part of the Gosling-Potter Internet Personality Project studying volunteers assessed via the Internet (Gosling et al., 2004; Srivastava et al., 2003).

2. Simulation studies

The majority of previous Monte Carlo studies that examined the role of sample size in factor analysis estimated factor recovery for a predefined range of sample sizes. In contrast, the present study estimated the minimum sample size that would yield a sample solution in good congruence with a population pattern (assuming a simple population pattern with common factors and equal loadings) for a combination of determinants (i.e., factor loadings, number of factors, and number of variables). The present study subsequently introduced a number of small distortions to a population pattern to investigate factor recovery in a realistic context.

2.1. Minimum sample size as a function of determinants

2.1.1. Method

The minimum sample size was estimated for population conditions with varying factor loadings ($\lambda = 0.2, 0.4, 0.6, 0.8, 0.9$), number of factors ($f = 1, 2, 3, 4, 8$), and number of variables ($p = 6, 12, 24, 48, 96$), except for $p < 2f$. The numerical ranges of the factors and variables were chosen to be representative for general factor analytical practice in psychological research (Henson & Roberts, 2006).

For each of the conditions under investigation, population solutions were defined to exhibit a simple pattern with equal loadings, as equal a number of loading variables per factor as possible, no secondary loadings, orthogonal factors, and no model error.[1] See Table 1 for an example.

The minimum sample size for each condition was estimated by means of a proportional controller. A Tucker's congruence coefficient (K) of 0.95 was considered the minimum threshold for "good agreement" (Lorenzo-Seva & Ten Berge, 2006).[2] The controller tuned N so that K converged to the 0.95 threshold. More precisely, the following procedure was repeated 5,000 times:

[1] This article defines simple structure as a special case of Thurstonian simple structure, also called independent cluster structure or ideal simple structure.

[2] Lorenzo-Seva and Ten Berge (2006) suggest the 0.95 threshold for good agreement on the basis of judgements of factor similarity by factor analytic experts. Note that others have used a 0.92 threshold for good and 0.98 for excellent agreement (MacCallum et al., 2001).

Table 1. Example of population pattern (λ = 0.8, f = 3, p = 24)

0.8	0.0	0.0
0.8	0.0	0.0
0.8	0.0	0.0
0.8	0.0	0.0
0.8	0.0	0.0
0.8	0.0	0.0
0.8	0.0	0.0
0.8	0.0	0.0
0.0	0.8	0.0
0.0	0.8	0.0
0.0	0.8	0.0
0.0	0.8	0.0
0.0	0.8	0.0
0.0	0.8	0.0
0.0	0.8	0.0
0.0	0.8	0.0
0.0	0.0	0.8
0.0	0.0	0.8
0.0	0.0	0.8
0.0	0.0	0.8
0.0	0.0	0.8
0.0	0.0	0.8
0.0	0.0	0.8
0.0	0.0	0.8

1. Based on the population solution, a sample observation matrix (N x p) was generated, using a method described by Hong (1999).
2. The Pearson correlation matrix of the sample observation matrix was submitted to principal axis factoring (maximum number of iterations: 9,999; iterative procedure continues until the maximum absolute difference of communalities was smaller than 10^{-3}) and oblique direct quartimin rotation (i.e., oblimin with gamma = 0) (Bernaards & Jennrich, 2005) by extracting f factors. To prevent optimism bias by screening solutions, unscreened data were used, that is, solutions that yielded Heywood cases (one or more variables with communalities equal to or higher than 1) were included in further analysis.
3. To recover the order and sign of the loadings, the Ks for the factor combinations (f x f) between the sample solution and the population solution were calculated.

Next, the reordering procedure of the sample solution started with the highest absolute *K* of the $f \times f$ calculated *K*s and proceeded towards the lowest *K* until the sign and order of all factors were recovered.

4. *K* was calculated between each reordered sample solution and the population pattern.
5. A new *N* was calculated as $N(i + 1) = N(i) - N(i)\cdot(K - 0.95)$, rounding away from $N(i)$. In other words, if $K > 0.95$, *N* was reduced, whereas, if $K < 0.95$, *N* was increased. Initial *N*, that is $N(1)$, was set at 1,000. A minimum *N* of 5 was set for controller stabilization. If *N* exceeded 10,000, the controlling phase was terminated and no estimated *N* was provided.

After the 5,000th repetition, the mean *N* of the last 4,500 repetitions was calculated, hereafter referred to as $N_{estimate}$. The first 500 iterations were omitted so that $N_{estimate}$ was based on the *N*s after the controller had stabilized.

The quality of $N_{estimate}$ was assessed during a verification phase. That is, for 5,000 new repetitions, median *K*, mean *K*, 5th percentile of *K*, the mean factor score correlation coefficient (*FSC*), and the proportion of sample solutions exhibiting one or more Heywood cases were calculated. The factor score correlation coefficient was inspired by the comparability coefficient described by Everett (1983). Bartlett factor scores based on the sample solution and Bartlett factor scores based on the population pattern were calculated. *FSC* was then defined as the correlation between the sample factor scores and the population factor scores, averaged over the *f* factors. Heywood variables were omitted when calculating the factor scores of the sample solution. Finally, Cohen's *d* effect size (*ES*) was calculated between the *f*-th and (*f+1*)-th eigenvalues of the unreduced correlation matrix as a descriptive measure of the size of their "gap".[3] An $ES = 4$ was considered a gap of adequate size, assuming two independent normally distributed eigenvalues with equal standard deviations and applying a threshold in the middle (i.e., $ES = 2$ from both means) implies that the correct number of factors can be identified in 95.5% of the solutions.

2.1.2. Results

The results in Table 2 show that factor recovery can be reliable with sample sizes well below 50. In agreement with the theoretical framework of MacCallum et al. (1999), lower sample sizes were needed when the level of loadings (λ) (therefore the

[3]) The *ES* index in this article was calculated from the eigenvalues of the unreduced correlation matrices (UCM, with 1s in the diagonal). It has been argued that it is more conceptually sensible to use the eigenvalues of the reduced correlation matrix (RCM, with communality estimates in the diagonal) when the goal is to identify the number of common factors (Fabrigar et al., 1999; Preacher & MacCallum, 2003). The present authors have repeated the subsampling study with *ES* based on the RCM (with communality estimates based on squared multiple correlations). Results showed that the difference in *ES* based on the UCM and the *ES* based on the RCM was always smaller than 10%, and that overall average *ES* was higher for the UCM, as compared to the RCM.

Table 2. Estimated N for satisfactory factor recovery for different factor loadings (λ), numbers of factors (f), and numbers of variables (p). For each condition, the median, mean, and 5th percentile (P5) of Tucker's congruence coefficient (K), the mean factor score correlation coefficient (*FSC*), the proportion of sample solutions exhibiting one or more Heywood cases, and Cohen's d effect size (*ES*) between the f-th and (f+1)-th eigenvalues are shown.

λ	f	p	N_{estimate}	Median K	Mean K	P5 K	Mean *FSC*	Heywood cases	*ES*
0.2	1	6	1524	0.961	0.950	0.879	0.952	0.000	6.23
		12	752	0.955	0.950	0.902	0.960	0.000	6.36
		24	470	0.953	0.951	0.919	0.970	0.000	7.53
		48	339	0.952	0.950	0.927	0.980	0.000	8.98
		96	274	0.951	0.950	0.933	0.987	0.000	10.43
	2	6	5849	0.958	0.950	0.883	0.949	0.000	6.78
		12	2571	0.954	0.950	0.916	0.955	0.000	6.91
		24	1438	0.952	0.950	0.927	0.962	0.000	8.38
		48	918	0.950	0.950	0.934	0.971	0.000	10.61
		96	676	0.950	0.950	0.939	0.981	0.000	13.47
	3	12	5363	0.953	0.950	0.914	0.952	0.000	6.88
		24	2829	0.951	0.950	0.931	0.959	0.000	8.33
		48	1732	0.950	0.950	0.937	0.967	0.000	11.16
		96	1197	0.950	0.950	0.942	0.976	0.000	14.96
	4	24	4602	0.950	0.950	0.932	0.956	0.000	7.94
		48	2750	0.950	0.950	0.939	0.964	0.000	11.24
		96	1827	0.950	0.950	0.942	0.973	0.000	15.57
	8	48	8695	0.950	0.950	0.942	0.957	0.000	10.18
		96	5390	0.950	0.950	0.945	0.964	0.000	15.75
0.4	1	6	102	0.963	0.950	0.871	0.954	0.004	5.31
		12	64	0.955	0.948	0.890	0.968	0.000	5.16
		24	52	0.953	0.949	0.905	0.982	0.000	5.66
		48	46	0.952	0.949	0.915	0.990	0.000	6.05
		96	44	0.952	0.950	0.919	0.995	0.000	6.46
	2	6	370	0.960	0.950	0.874	0.946	0.015	6.61
		12	186	0.954	0.950	0.911	0.963	0.000	6.16
		24	134	0.951	0.949	0.924	0.976	0.000	7.02
		48	112	0.951	0.950	0.931	0.986	0.000	8.21
		96	101	0.950	0.950	0.936	0.992	0.000	9.26

(*table continues*)

Table 2. (continued)

λ	*f*	*p*	N_{estimate}	Median	Mean	P5	Mean	Heywood	*ES*
				K	*K*	*K*	*FSC*	cases	
	3	6	1159	0.953	0.949	0.888	0.938	0.001	8.84
		12	353	0.954	0.950	0.916	0.958	0.001	6.27
		24	234	0.951	0.950	0.929	0.972	0.000	7.48
		48	186	0.951	0.950	0.936	0.983	0.000	9.28
		96	163	0.950	0.950	0.939	0.990	0.000	10.80
	4	12	589	0.954	0.950	0.913	0.953	0.006	6.40
		24	349	0.951	0.950	0.932	0.969	0.000	7.28
		48	270	0.950	0.950	0.938	0.980	0.000	9.63
		96	230	0.950	0.950	0.941	0.988	0.000	11.94
	8	24	977	0.951	0.950	0.934	0.958	0.001	6.53
		48	678	0.950	0.950	0.942	0.972	0.000	9.34
		96	541	0.950	0.950	0.944	0.982	0.000	13.62
0.6	1	6	18	0.965	0.940	0.813	0.946	0.046	4.42
		12	15	0.961	0.943	0.844	0.969	0.004	4.39
		24	13	0.955	0.940	0.840	0.982	0.001	4.46
		48	12	0.952	0.938	0.847	0.991	0.000	4.61
		96	12	0.951	0.940	0.860	0.995	0.000	4.69
	2	6	59	0.960	0.950	0.877	0.945	0.120	5.35
		12	39	0.955	0.948	0.896	0.968	0.001	5.15
		24	34	0.952	0.948	0.913	0.983	0.000	5.66
		48	31	0.951	0.948	0.920	0.991	0.000	6.19
		96	30	0.951	0.948	0.923	0.995	0.000	6.43
	3	6	208	0.950	0.949	0.903	0.913	0.404	8.42
		12	67	0.954	0.949	0.910	0.964	0.009	5.26
		24	55	0.951	0.949	0.924	0.981	0.000	5.94
		48	50	0.949	0.948	0.929	0.989	0.000	6.77
		96	49	0.951	0.950	0.934	0.994	0.000	7.72
	4	12	99	0.952	0.949	0.911	0.956	0.051	5.20
		24	78	0.951	0.950	0.929	0.978	0.000	5.97
		48	71	0.950	0.949	0.935	0.988	0.000	7.31
		96	68	0.950	0.950	0.938	0.993	0.000	8.56
	8	24	179	0.951	0.950	0.933	0.967	0.004	5.36
		48	156	0.950	0.950	0.941	0.983	0.000	7.59
		96	146	0.950	0.950	0.943	0.990	0.000	10.15

(*table continues*)

Table 2. (continued)

λ	*f*	*p*	$N_{estimate}$	Median	Mean	P5	Mean	Heywood	*ES*
				K	*K*	*K*	*FSC*	cases	
0.8	1	6	6	0.984	0.935	0.686	0.955	0.331	4.95
		12	6	0.983	0.948	0.786	0.975	0.176	4.93
		24	6	0.982	0.954	0.822	0.987	0.099	5.06
		48	6	0.982	0.957	0.840	0.994	0.048	5.24
		96	6	0.981	0.959	0.850	0.997	0.025	5.24
	2	6	12	0.960	0.941	0.815	0.948	0.542	3.75
		12	11	0.956	0.940	0.843	0.972	0.134	3.85
		24	10	0.949	0.937	0.856	0.983	0.044	3.80
		48	10	0.949	0.940	0.876	0.990	0.008	4.07
		96	10	0.949	0.941	0.882	0.993	0.002	4.32
	3	6	21	0.958	0.949	0.886	0.901	0.845	3.73
		12	17	0.952	0.942	0.873	0.969	0.181	3.64
		24	17	0.952	0.947	0.907	0.985	0.011	4.08
		48	17	0.952	0.949	0.916	0.992	0.000	4.53
		96	17	0.952	0.949	0.922	0.995	0.000	4.84
	4	12	24	0.954	0.948	0.900	0.967	0.325	3.61
		24	23	0.952	0.948	0.919	0.984	0.010	4.11
		48	23	0.951	0.949	0.927	0.991	0.000	4.72
		96	23	0.950	0.949	0.929	0.994	0.000	5.15
	8	24	45	0.950	0.949	0.927	0.977	0.105	3.71
		48	47	0.951	0.950	0.938	0.989	0.000	5.10
		96	47	0.950	0.950	0.939	0.993	0.000	6.31
0.9	1	6	5	0.996	0.961	0.837	0.978	0.442	7.19
		12	5	0.996	0.978	0.895	0.989	0.338	7.07
		24	5	0.996	0.978	0.901	0.994	0.247	7.00
		48	5	0.996	0.981	0.914	0.997	0.183	7.11
		96	5	0.995	0.978	0.905	0.998	0.135	7.22
	2	6	7	0.974	0.951	0.816	0.968	0.771	3.37
		12	6	0.958	0.931	0.748	0.976	0.704	3.07
		24	6	0.955	0.934	0.806	0.984	0.567	3.17
		48	6	0.954	0.935	0.814	0.988	0.457	3.19
		96	6	0.954	0.937	0.835	0.991	0.344	3.24

(*table continues*)

Table 2. (continued)

λ	*f*	*p*	N_{estimate}	Median	Mean	P5	Mean	Heywood	*ES*
				K	*K*	*K*	*FSC*	cases	
	3	6	8	0.954	0.929	0.769	0.896	0.938	2.56
		12	9	0.955	0.939	0.838	0.975	0.660	2.83
		24	9	0.952	0.940	0.867	0.985	0.350	2.91
		48	9	0.950	0.941	0.878	0.989	0.149	3.05
		96	9	0.949	0.941	0.886	0.991	0.064	3.17
	4	12	12	0.956	0.944	0.861	0.974	0.772	2.72
		24	12	0.951	0.943	0.887	0.985	0.277	2.92
		48	12	0.949	0.945	0.904	0.990	0.061	3.14
		96	12	0.948	0.944	0.908	0.992	0.016	3.32
	8	24	23	0.953	0.947	0.919	0.982	0.532	2.84
		48	24	0.950	0.949	0.930	0.990	0.017	3.50
		96	25	0.951	0.951	0.936	0.993	0.000	4.11

Note. All solutions were based on 5,000 repetitions.

communalities) was high, the number of factors (*f*) small, and the number of variables (*p*) high. For loadings higher than 0.8 and one factor, even sample sizes smaller than 10 were sufficient for factor recovery. The level of loadings was a very strong determinant. For example, when loadings were as high as 0.9, and even with a high number of factors ($f = 4$) and a limited number of variables ($p = 12$), a sample size of 12 sufficed.

A larger number of variables improved factor recovery, particularly when loadings were low. No practical objection for performing EFA was found in conditions where the number of variables exceeded the sample size. In fact, increasing the number of variables reduced the minimum *N*, also when $p > N$.

Table 2 shows that for constant mean *K*, *ES* was lowest when high λ was combined with low *p* and low *N*, indicating that researchers should be cautious when deciding on the number of factors, particularly under such circumstances. In most conditions, however, *ES* was greater than 4. The highest *ES* was found in patterns with low λ, high *p*, and high *N*. Increasing *p* was beneficial for *ES*, even when *N* was decreased.

Table 2 also reveals the different tendencies of the factor recovery indices. Although the mean/median *K* was kept constant at 0.95, the 5th percentile of *K* systematically increased with an increase of *p*, signifying a favourable distributional change of *K*. *FSC* consistently and strongly improved with an increase of *p*, profiting from the additional information provided by extra variables. The proportion of sample solutions exhibiting Heywood cases reduced with higher *p,* whereas it increased for higher λ. This phenomenon can be attributed to the fact that increased λ elevates the risk of communalities higher than 1, due to sampling error. Note that the presence of Heywood cases was not detrimental to the recovery of the population pattern per se, as high *K* and high *FSC* could still be obtained in the unscreened solutions.

A more detailed analysis was conducted to gain insight into the interactions between the determinants. Mean *K,* mean *FSC*, and *ES* were calculated for a wide range of *f* (between 1 and 8) and *p* (logarithmically spaced between 10 and 200), except for $p < 2f$, for six combinations of sample sizes (small: $N = 25$, medium: $N = 100$, and high: $N = 1{,}000$) and levels of loadings (low: $\lambda = 0.4$ and high: $\lambda = 0.9$). The results are shown in Figure 1. Increasing *N* was always beneficial. Also apparent is that *f* had a relatively strong influence, with mean *K* and mean *FSC* reducing when *f* increased. The effect of *p*, on the one hand, depended on λ: For low λ, increasing *p* resulted in higher mean *K* and *ES*, whereas, for high λ, increasing *p* had a much smaller positive effect. For *FSC*, on the other hand, a higher *p* was beneficial for both high and low λ. These findings agree with the theoretical framework presented by MacCallum et al. (1999) demonstrating that a high *p*/*f* is advantageous to factor recovery and that this effect diminishes with increasing λ. However, the present results also showed that *p*/*f* is not a comprehensive measure, as *p* and *f* have clearly distinct effects on factor recovery.

2.2. The role of distortions

2.2.1. Method

In reality, models rarely exhibit a perfectly simple structure. Moreover, models are imperfect, leaving a part of reality undepicted. For this reason, we systematically evaluated the role of various distortions (divided into 13 groups of four conditions each) in a baseline population solution with a small *N* but large λ, low *f*, and large *p* ($N = 17$, $\lambda = 0.8$, $f = 3$, $p = 24$). The corresponding pattern solution is shown in Table 1. Iterative principal factor analysis was performed with oblimin rotation for all the investigated groups of distortions. Sufficient repetitions were performed for each condition so that the 95% confidence interval of the mean *K* was narrower than 0.001. The design of the simulation is summarized in Table 3. As an example, Table 4 shows the first population pattern of each investigated group.

Group 1. Interfactor correlation (ifc = 0.1, 0.3, 0.5, 0.7) for one pair of factors
This group examined the effect of various levels of *ifc* between two factors.

Group 2. Interfactor correlation (ifc = 0.1, 0.3, 0.5, 0.7) between all factors
Same as Group 1, but here all three combinations of factors were correlated, providing a more severe test case.

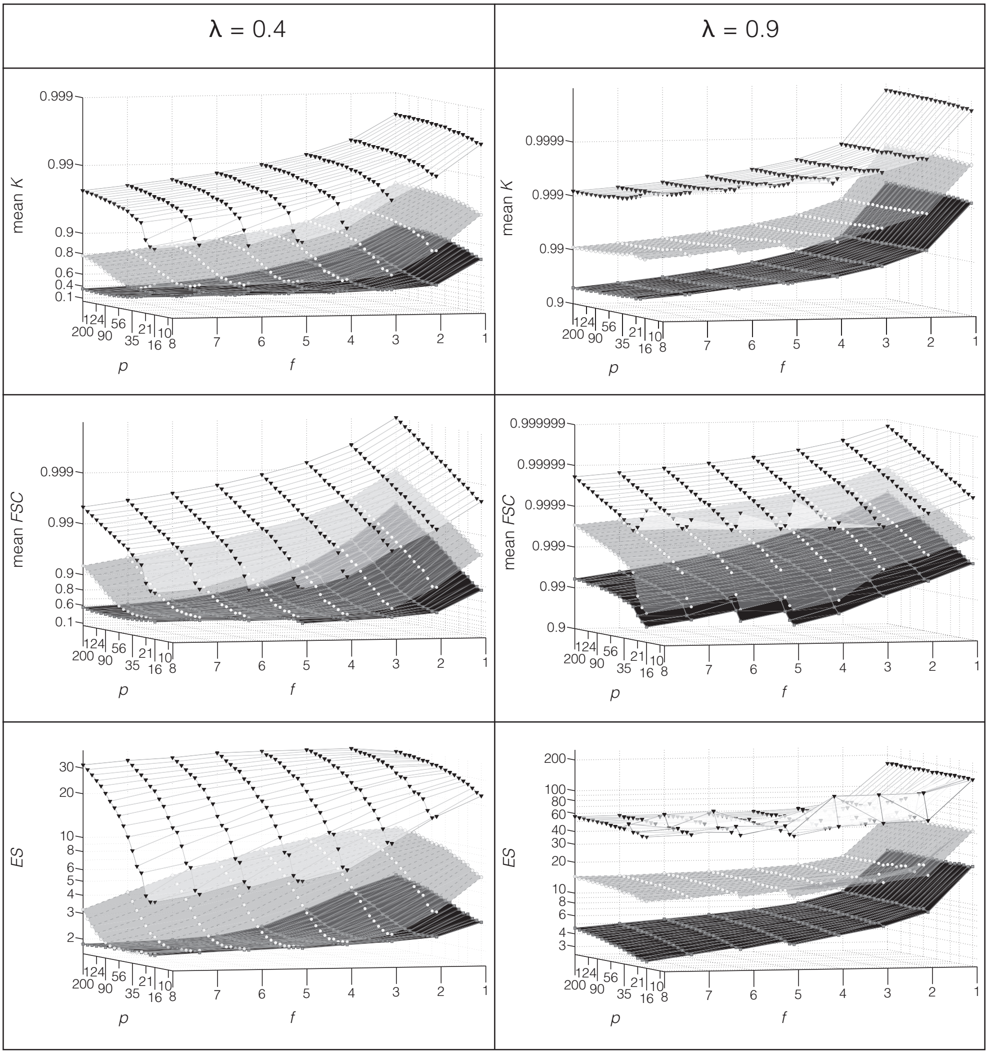


Figure 1. Main effects and interactions among the determinants of factor analysis. The plots show the factor recovery indices (mean Tucker's congruence coefficient (*K*), mean factor score correlation coefficient (*FSC*), and the Cohen's *d* effect size (*ES*) between the *f*-th and (*f* +1)-th eigenvalues) as function of a wide range of *f* and *p,* and two levels of loadings (λ = 0.4 and λ = 0.9). Three levels of sample size are shown in each plot, that is, *N* = 25 (white), 100 (grey), and 1,000 (black).

Group 3. Model error, altering the amount of variance

To investigate whether model error plays a role in factor recovery for small *N*, random model error was introduced for every repetition by means of 200 minor factors

Table 3. Design of the simulation study investigating the role of distortions

Group 1. Interfactor correlation for one pair of factors	Group 8. Unequal loadings between factors
small (0.1)	small deviations (0.85/0.8/0.75)
medium (0.3)	medium deviations (0.9/0.8/0.69)
large (0.5)	large deviations (0.95/0.8/0.61)
very large (0.7)	very large deviations (0.99/0.8/0.55)
Group 2. Interfactor correlation between all factors	**Group 9. Unequal loadings within factors**
small (0.1)	small deviations (0.85/0.75)
medium (0.3)	medium deviations (0.9/0.69)
large (0.5)	large deviations (0.95/0.61)
very large (0.7)	very large deviations (0.99/0.55)
Group 3. Model error; altering the amount of variance [a]	**Group 10. Secondary loadings**
small (0.05)	2 variables, loadings 0.2/-0.2
medium (0.10)	2 variables, loadings 0.4/-0.4
large (0.15)	4 variables, loadings 0.2/-0.2
very large (0.20)	4 variables, loadings 0.4/-0.4
Group 4. Model error; altering the distribution of minor factors [a]	**Group 11. Random distortions of all loadings [a]**
ε = small (0.05)	small (range 0.05)
ε = medium (0.15)	medium (range 0.10)
ε = large (0.25)	large (range 0.15)
ε = very large (0.35)	very large (range 0.20)
Group 5. Low loadings (0.6) added	**Group 12. Unequal *p*/*f* (one weak factor) [b]**
adding 6 variables	*p*/*f* = 8, 8, 6
adding 12 variables	*p*/*f* = 8, 8, 4
adding 24 variables	*p*/*f* = 8, 8, 3
adding 96 variables	*p*/*f* = 8, 8, 2
Group 6. Low loadings (0.6) replacing high loadings	**Group 13. Unequal *p*/*f* (two weak factors) [b]**
replacing 3 variables	*p*/*f* = 8, 6, 6
replacing 6 variables	*p*/*f* = 8, 4, 4
replacing 12 variables	*p*/*f* = 8, 3, 3
replacing 18 variables	*p*/*f* = 8, 2, 2
Group 7. Altering the number of variables	
p = 12	
p = 15	
p = 18	
p = 48	

[a] A different population pattern was produced for each repetition for all conditions of groups 3, 4, and 11.

[b] The numbers refer to the number of variables per factor with a 0.8 loading.

Table 4. Population patterns used in the simulations

Group 5			Group 6			Group 7			Group 8			Group 9		
0.8	0	0	0.8	0	0	0.8	0	0	0.85	0	0	0.85	0	0
0.8	0	0	0.8	0	0	0.8	0	0	0.85	0	0	0.747	0	0
0.8	0	0	0.8	0	0	0.8	0	0	0.85	0	0	0.85	0	0
0.8	0	0	0.8	0	0	0.8	0	0	0.85	0	0	0.747	0	0
0.8	0	0	0.8	0	0	0	0.8	0	0.85	0	0	0.85	0	0
0.8	0	0	0.8	0	0	0	0.8	0	0.85	0	0	0.747	0	0
0.8	0	0	0.8	0	0	0	0.8	0	0.85	0	0	0.85	0	0
0.8	0	0	0.6	0	0	0	0.8	0	0.85	0	0	0.747	0	0
0	0.8	0	0	0.8	0	0	0	0.8	0	0.8	0	0	0.85	0
0	0.8	0	0	0.8	0	0	0	0.8	0	0.8	0	0	0.747	0
0	0.8	0	0	0.8	0	0	0	0.8	0	0.8	0	0	0.85	0
0	0.8	0	0	0.8	0	0	0	0.8	0	0.8	0	0	0.747	0
0	0.8	0	0	0.8	0				0	0.8	0	0	0.85	0
0	0.8	0	0	0.8	0				0	0.8	0	0	0.747	0
0	0.8	0	0	0.8	0				0	0.8	0	0	0.85	0
0	0.8	0	0	0.6	0				0	0.8	0	0	0.747	0
0	0	0.8	0	0	0.8				0	0	0.747	0	0	0.85
0	0	0.8	0	0	0.8				0	0	0.747	0	0	0.747
0	0	0.8	0	0	0.8				0	0	0.747	0	0	0.85
0	0	0.8	0	0	0.8				0	0	0.747	0	0	0.747
0	0	0.8	0	0	0.8				0	0	0.747	0	0	0.85
0	0	0.8	0	0	0.8				0	0	0.747	0	0	0.747
0	0	0.8	0	0	0.8				0	0	0.747	0	0	0.85
0	0	0.8	0	0	0.6				0	0	0.747	0	0	0.747
0.6	0	0												
0.6	0	0												
0	0.6	0												
0	0.6	0												
0	0	0.6												
0	0	0.6												

(*table continues*)

Table 4. (continued)

Group 10			Group 11			Group 12			Group 13		
0.8	0.2	0	0.791	-0.019	-0.011	0.8	0	0	0.8	0	0
0.8	-0.2	0	0.807	0.002	0.004	0.8	0	0	0.8	0	0
0.8	0	0	0.805	0.017	-0.012	0.8	0	0	0.8	0	0
0.8	0	0	0.778	-0.013	0.014	0.8	0	0	0.8	0	0
0.8	0	0	0.819	-0.024	-0.018	0.8	0	0	0.8	0	0
0.8	0	0	0.811	-0.018	0.014	0.8	0	0	0.8	0	0
0.8	0	0	0.776	-0.001	0.001	0.8	0	0	0.8	0	0
0.8	0	0	0.812	0.019	0.004	0.8	0	0	0.8	0	0
0	0.8	0	-0.010	0.818	0.001	0	0.8	0	0	0.8	0
0	0.8	0	0.013	0.825	-0.018	0	0.8	0	0	0.8	0
0	0.8	0	-0.020	0.813	-0.009	0	0.8	0	0	0.8	0
0	0.8	0	-0.014	0.819	0.018	0	0.8	0	0	0.8	0
0	0.8	0	-0.001	0.793	-0.003	0	0.8	0	0	0.8	0
0	0.8	0	0.019	0.811	-0.022	0	0.8	0	0	0.8	0
0	0.8	0	0.004	0.789	0.019	0	0.8	0	0	0	0.8
0	0.8	0	-0.004	0.815	-0.012	0	0.8	0	0	0	0.8
0	0	0.8	0.016	-0.015	0.789	0	0	0.8	0	0	0.8
0	0	0.8	0.010	-0.004	0.786	0	0	0.8	0	0	0.8
0	0	0.8	-0.003	0.006	0.805	0	0	0.8	0	0	0.8
0	0	0.8	-0.002	0.003	0.803	0	0	0.8	0	0	0.8
0	0	0.8	0.019	-0.013	0.818	0	0	0.8			
0	0	0.8	0.007	0.008	0.786	0	0	0.8			
0	0	0.8	-0.018	-0.018	0.790						
0	0	0.8	-0.014	0.001	0.804						

Note. The first condition described in the text is shown for each group.

explaining (a) 0.05, (b) 0.1, (c) 0.15, and (d) 0.2 of the variance. The data generation parameter determining the distribution of successive minor factors was set at ε = 0.1.

Group 4. Model error, altering the distribution of the minor factors
In this group, the distribution of minor factors explaining 0.2 of the variance was altered by varying ε from 0.05 to 0.15, 0.25, and 0.35. Larger values of ε causes the contribution of the minor factors to be more skewed in favour of the earlier factors in the sequence (MacCallum & Tucker, 1991).

Group 5. Low loadings (0.6) added
The simulation results in Table 2 on the factor analytic determinants showed that the addition of variables improved factor recovery. Adding low loading variables, however, inevitably decreases average communalities amongst all variables. Considering that both communalities and the number of variables are important determinants of factor recovery, the question is whether the addition of variables improves factor recovery, even when the added variables reduce average communalities. For this reason, the behaviour of population patterns with a number (6, 12, 24, 96) of extra variables with low loadings (0.6) was studied.[4]

Group 6. Low loadings (0.6) replacing high loadings
A number (3, 6, 12, 18) of high loading variables were replaced with low loading (0.6) variables. This was expected to cause a stronger distortion than Group 5 because the number of high loading variables was also reduced.

Group 7. Altering the number of variables
The number of variables was altered from 24 to 12, 15, 18, and 48, in order to investigate whether a discontinuity appears in factor recovery when factor analysing samples in which the number of variables exceeds the sample size.

Group 8. Unequal loadings between factors
The level of the loadings among the three factors was varied in such a way that the average communalities of all variables remained equal to the baseline condition (i.e., 0.64). The following four combinations were investigated: (a) 0.85/0.8/0.75 ($= [0.8^2 - (0.85^2 - 0.8^2)]^{0.5}$), (b) 0.9/0.8/0.69, (c) 0.95/0.8/0.61, and (d) 0.99/0.8/0.55.

[4]) A loading of 0.6 was considered as low for the sample size (N = 17) under investigation. This was based on the findings of the first part of the simulations (Table 2): for λ = 0.6, f = 3, p = 24, the required minimum N for good agreement (K = 0.95) was 55.

Group 9. Unequal loadings within factors
The loadings within each of the three factors were alternated in such a way that the average communalities were equal to those in the baseline condition. The following four combinations of alternate nonzero loadings were investigated: (a) 0.85/0.75 ($= [0.8^2 - (0.85^2 - 0.8^2)]^{0.5}$), (b) 0.9/0.69, (c) 0.95/0.61, and (d) 0.99/0.55.

Group 10. Secondary loadings
The effect of adding two or four secondary loadings of low (0.2) as well as high (0.4) level was examined. Alternating signs of the secondary loadings were used to prevent rotation toward a different solution.

Group 11. Random distortions of all loadings
In reality, population patterns are not homogeneous. Therefore, random distortions of all loadings were introduced. More precisely, four levels of uniform random loadings (ranges 0.05, 0.1, 0.15, and 0.2) were added to the baseline.

Group 12. Unequal p/f (one weak factor)
Equal *p/f* rarely occurs in reality. Therefore the third factor was weakened by decreasing the number of variables that loaded on this factor.

Group 13. Unequal p/f (two weak factors)
Factor 2 and 3 were weakened by decreasing the number of variables that loaded on these factors. This means tested the impact of weakening two out of three factors on factor recovery.

2.2.2. Results

Figure 2 shows the mean *K*, mean *FSC*, the proportion of sample solutions exhibiting Heywood cases, and *ES* for each of the 13 groups.

Groups 1–2. Interfactor correlation
When one pair of factors was very strongly (0.7) correlated or all factors were strongly (0.5) correlated, mean *K* and mean *FSC* deteriorated considerably. Small interfactor correlation (up to 0.3) disturbed *K* and *FSC* to a far lesser extent. The proportion of sample solutions exhibiting Heywood cases increased slightly when all three factors were strongly correlated. Correlated factors negatively affected *ES* more than any other distortion did.

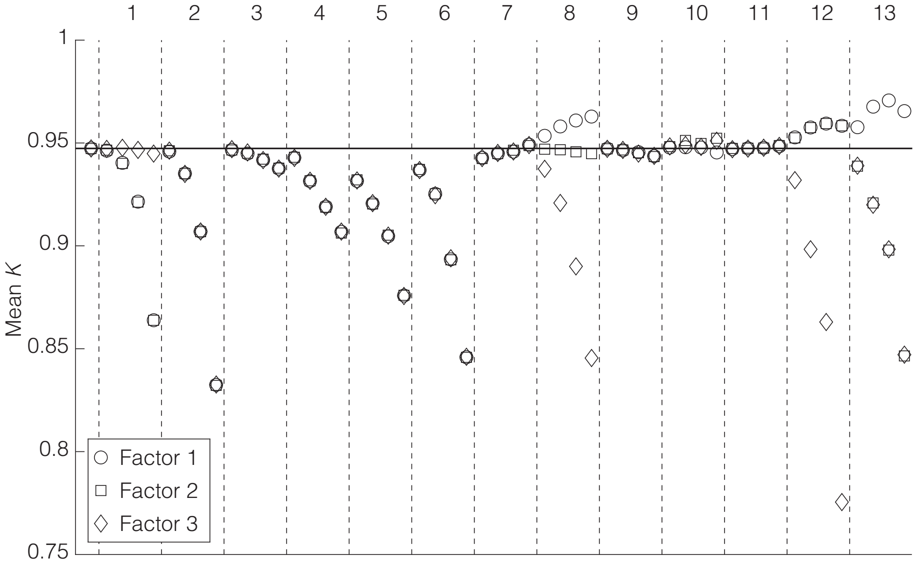
Groups 3–4. Model error
Model error slightly worsened factor recovery. This effect was seen in all four indices. Introducing a large model error (0.2) across a more skewed distribution of

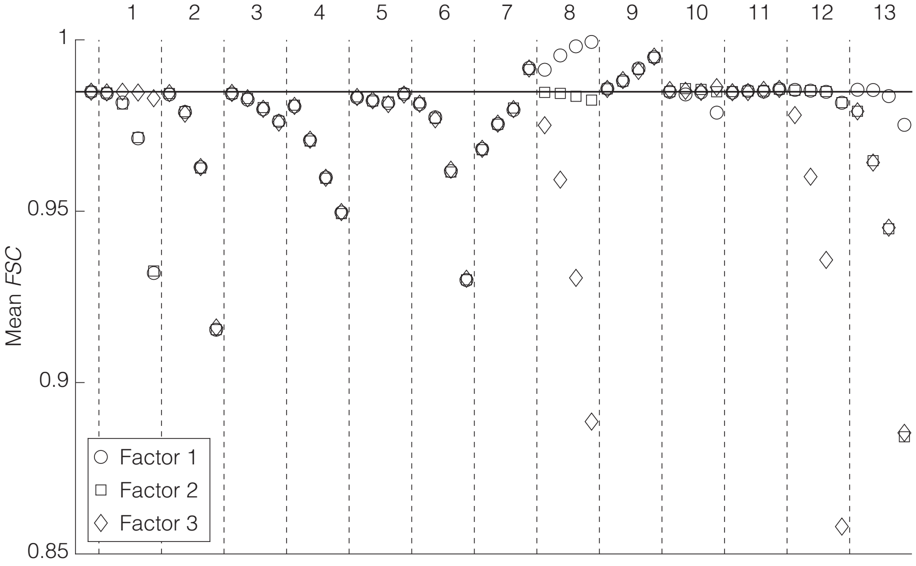
minor factors (ε = 0.25 or 0.35) caused a relatively strong degradation of factor recovery as compared to the effect of a less skewed distribution.

Groups 5–6. Low loadings (0.6) added or replacing high loadings
Adding low loading variables worsened mean *K*. Mean *FSC* slightly decreased for a small number of added low loadings but recovered for a larger (96) number of added low loadings. This can be explained by the fact that *K* takes into account all (low as well as high) loadings, whereas *FSC* is based on factor scores, obtaining the information from all the manifest variables. In other words, *FSC* benefited from (or, at least, stayed unaffected by) additional information. On the other hand, as an index of factor loadings similarity, *K* was more easily disturbed by the presence of variables of low quality. The deterioration of *K* was even more dramatic when low loading variables replaced high loading variables, whereas *FSC* degraded mainly when 18 out of the 24 variables had been replaced by low loadings. The proportion of sample solutions exhibiting Heywood cases reduced when low loading variables were added but increased when low loading variables replaced high loading variables. Appending low loadings influenced *ES* only slightly. However, this index degraded, when low loadings replaced high loadings.

Group 7. Altering the number of variables
An increased number of variables slightly improved mean *K*, considerably improved mean *FSC*, and strongly suppressed the proportion of sample solutions exhibiting Heywood cases. In an additional test, driven more by theoretical interest rather than a realistic approach, factor recovery was estimated for 600 variables, a level at which mean *FSC* reached near unity (0.997). In accordance with the first series of simulations, increasing the number of variables increased *ES*. The results of this group also showed that there was no discontinuity whatsoever with respect to factor recovery at the point where the number of variables exceeded the sample size.

Groups 8–9. Unequal loadings between or within factors
For unequal loadings between factors (Group 8), the recovery of factors with high loadings improved (with the *FSC* of the first factor reaching 0.999 in the fourth condition), whereas the recovery of factors with low loadings deteriorated. Unequal loadings within factors (Group 9) strongly increased the likelihood of Heywood cases. However, this did not deteriorate factor recovery: mean *K* remained constant while mean *FSC* increased up to near unity (0.995) in the fourth condition, which had a very large proportion of Heywood cases. *ES* decreased with unequal loadings between factors (Group 8) but was less sensitive to unequal loadings within factors (Group 9).


1
2
3
4
5
6
7
8
9
10
11
12
13
1
0.95
0.9
0.85
0.8
0.75
Mean *K*
○ Factor 1
□ Factor 2
◇ Factor 3


1
2
3
4
5
6
7
8
9
10
11
12
13
1
0.95
0.9
0.85
Mean *FSC*
○ Factor 1
□ Factor 2
◇ Factor 3

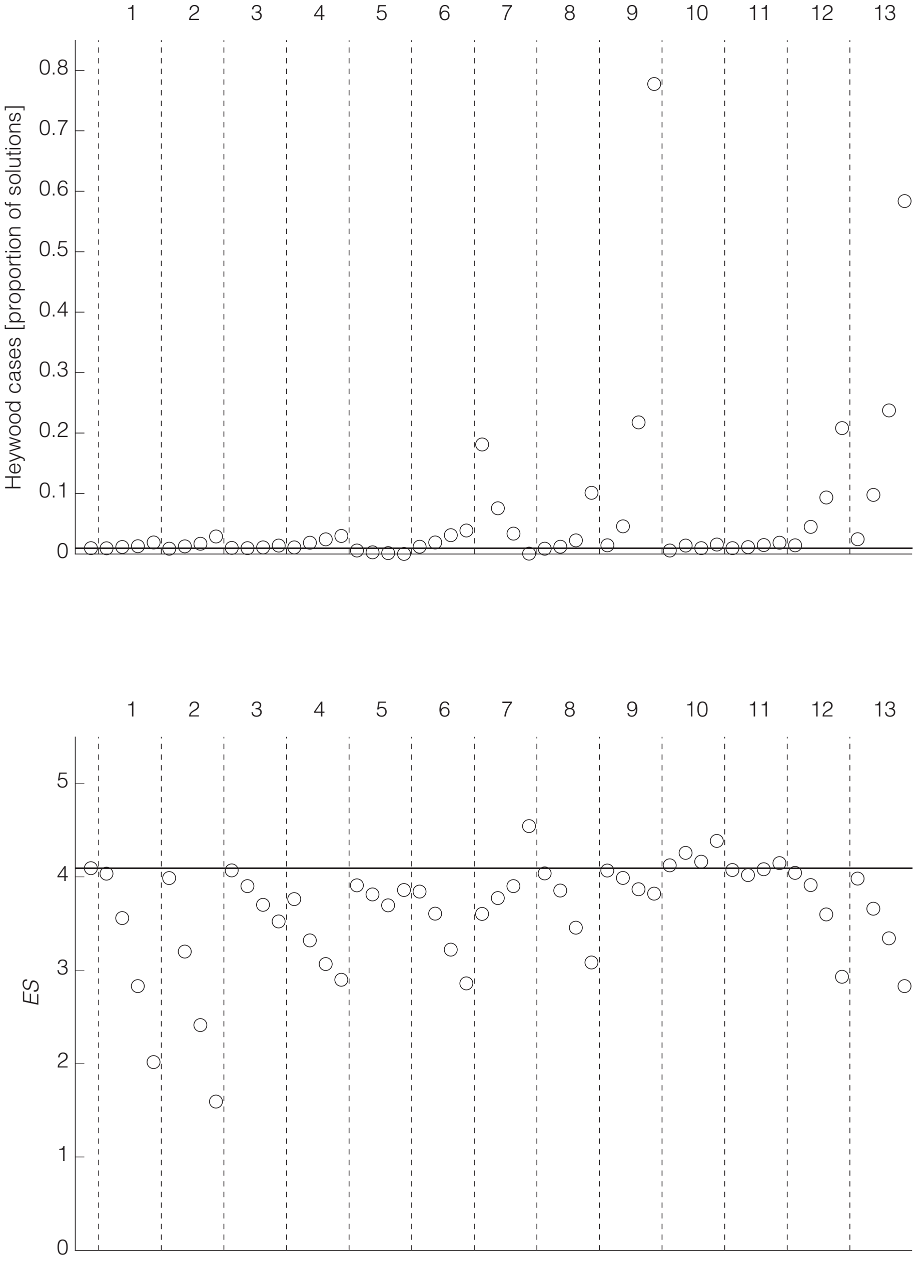


Figure 2. Factor recovery for the 13 investigated groups. Left top: mean Tucker's congruence coefficient *K*, left bottom: mean factor score correlation coefficient *FSC*, right top: proportion of sample solutions exhibiting one or more Heywood cases, and right bottom: the Cohen's *d* effect size (*ES*) between the third and fourth eigenvalues. The horizontal line represents the average factor recovery for the baseline condition.

Group 10. Secondary loadings

Secondary loadings hardly influenced factor recovery. Only when the number and level of secondary loadings were the highest tested did mean *FSC* slightly decrease. Mean *K*, on the other hand, slightly increased when secondary loadings were high (0.4). Moreover, secondary loadings were beneficial to *ES*.

Group 11. Random distortions of all loadings

Randomly distorted loadings hardly influenced factor recovery, signifying that the positive effect of the presence of high loadings compensated for the negative effect of random low loadings.

Groups 12–13. Unequal p/f (one or two weak factors)

Low p/f had a negative effect on the recovery of the corresponding factor. The mean *FSC* of the worst condition ($p/f = 2$) was still higher than 0.85, so even for a very weak factor, factor recovery was not necessarily grossly wrong in a small *N* scenario. The recovery of the strong factors improved in terms of *K*. A low p/f increased the proportion of sample solutions exhibiting Heywood cases and weakened *ES*, diminishing the odds of correctly estimating the number of factors.

Group summary

The investigated baseline ($N = 17$, $\lambda = 0.8$, $f = 3$, $p = 24$) was noticeably robust against single small distortions. Each of the indices (mean *K*, mean *FSC,* Heywood cases, and *ES*) was sensitive to different distortions. The most serious degradation of factor recovery was caused by a highly unequal distribution of variables between factors (unequal p/f). In addition, *ES* was highly sensitive to interfactor correlations. Replacing high with low loadings or having unequal loadings between factors also negatively influenced *ES*.

3. Subsampling study

A subsampling study was carried out to investigate whether the findings of the simulation study are realistic and hold for actual data. An empirical dataset with a large sample size was used, acting as a population. The dataset consisted of 280,691 participants (mean age = 30.4, SD = 8.5 years, 54% women) who filled in the 44-item Big Five Inventory (Gosling et al., 2004; Srivastava et al., 2003). All selected participants indicated that they were filling in the inventory for the first time. Only participants who filled in the English version of the inventory, answering all items without giving identical answers to all 44 items were included. Subsamples were drawn from the population sample, and factor recovery was assessed between the subsampling and the population solution.

Table 5. Mean, minimum, and maximum of primary loadings, secondary loadings, communalities, and interfactor correlations of the population solutions for the investigated number of factors

f [a]	*p*	mean of primary loadings (min/max)	mean of secondary loadings (min/max)	mean communalities (min/max)	mean interfactor correlation (min/max)
1 (E)	8	0.64(0.50/0.77)	–	0.42(0.25/0.59)	–
2 (E,O)	18	0.58(0.23/0.79)	0.10(0.02/0.28)	0.37(0.08/0.60)	0.17(0.17/0.17)
3 (E,O,C)	27	0.58(0.24/0.80)	0.11(0.03/0.27)	0.38(0.08/0.62)	0.14(0.12/0.16)
4 (E,O,C,N)	35	0.59(0.23/0.81)	0.13(0.05/0.28)	0.40(0.10/0.63)	0.14(0.05/0.24)
5 (E,O,C,N,A)	44	0.57(0.24/0.80)	0.15(0.07/0.27)	0.39(0.10/0.62)	0.12(0.01/0.20)

Note. All numbers are based on the absolute values of the pattern matrix and absolute values of the interfactor correlations.

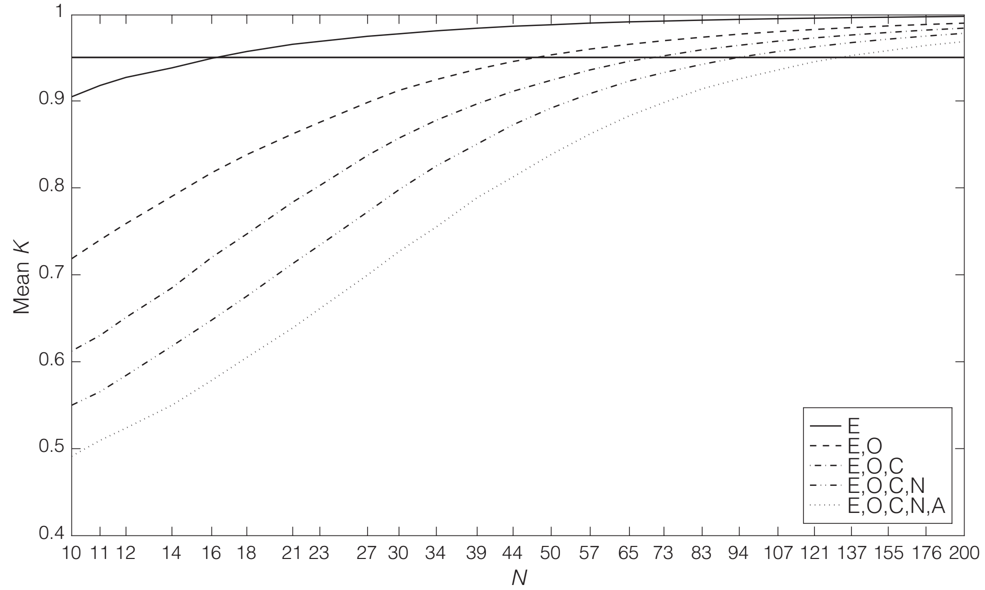
[a] E = extraversion, O = openness, C = conscientiousness, N = neuroticism, A = agreeableness.

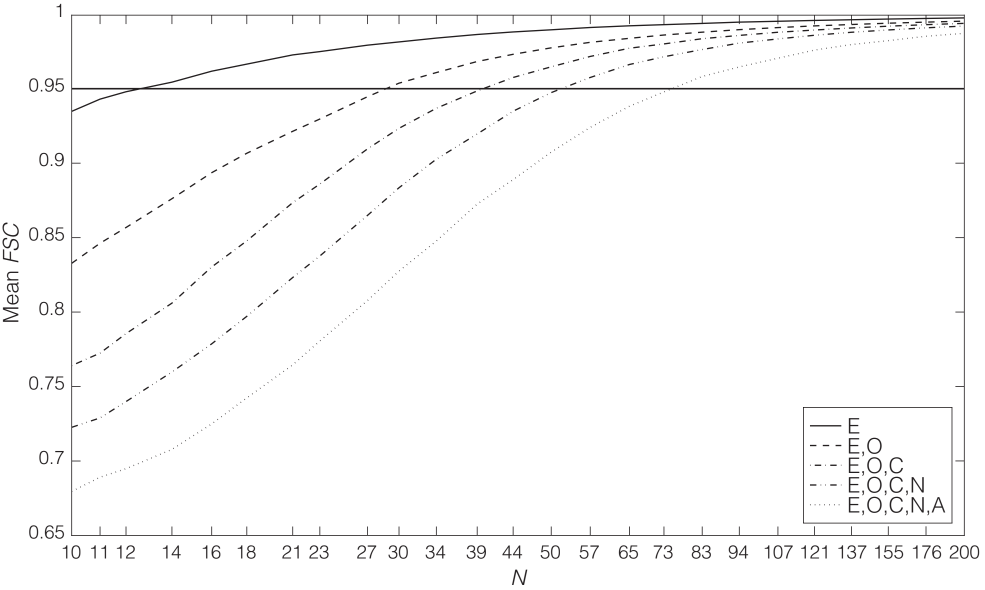
3.1. Method

Factor recovery was estimated for a range of 25 subsample sizes spaced logarithmically between 10 and 200. For each subsample size, 10,000 random subsamples were drawn from the population and factor analysed as in the simulation study. To investigate the role of the number of factors as a determinant of factor analytic performance, factor recovery was assessed when retaining from 1 up to 5 factors. Variables were selected so that each factor contained the 8 to 10 variables representing the corresponding personality trait. Table 5 summarizes the results of the five population patterns. Communalities were wide. Interfactor correlations were low, so these should hardly affect factor recovery, according to the simulations. Factor recovery was evaluated using mean *K*, mean *FSC*, the proportion of solutions exhibiting a Heywood case, and *ES*.

3.2. Results

Figure 3 shows the factor recovery results. For one extracted factor, a sample size around 13 and 17 was adequate for satisfactory *FSC* and *K* (= 0.95), whereas $f = 2$ required a sample size between 30 (for *FSC* = 0.95) and 50 (for *K* = 0.95). When retaining all factors of the Big Five ($f = 5$), a considerably larger sample size (80 to 140) was needed. For all numbers of factors, the proportion of solutions exhibiting a Heywood case was below 0.05 for sample sizes greater than 17. For one extracted factor, a sample size of 10 was sufficient for *ES* = 4. In contrast, when all 5 factors were retained, a larger sample size (140) was required to guarantee an adequate *ES*.


Mean K
1
0.9
0.8
0.7
0.6
0.5
0.4
10 11 12 14 16 18 21 23 27 30 34 39 44 50 57 65 73 83 94 107 121 137 155 176 200
N
E
E,O
E,O,C
E,O,C,N
E,O,C,N,A


Mean FSC
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
10 11 12 14 16 18 21 23 27 30 34 39 44 50 57 65 73 83 94 107 121 137 155 176 200
N
E
E,O
E,O,C
E,O,C,N
E,O,C,N,A

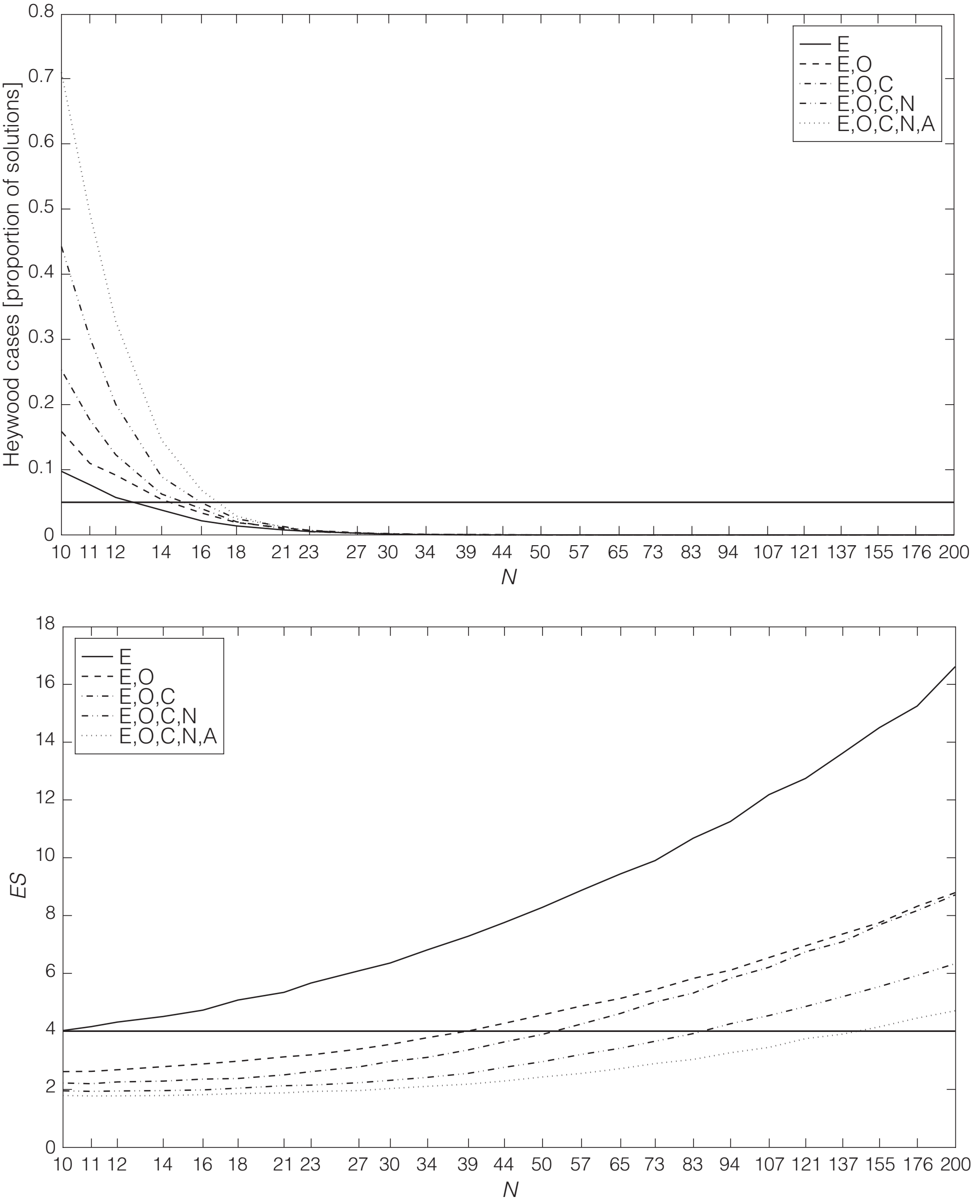


Figure 3. Subsampling factor recovery for the Big Five Inventory. Left top: mean Tucker's congruence coefficient (*K*), left bottom: mean factor score correlation coefficient (*FSC*), right top: proportion of sample solutions exhibiting one or more Heywood cases, and right bottom: the Cohen's *d* effect size (*ES*) between the *f*-th and (*f* +1)-th eigenvalues. The horizontal line represents a threshold for satisfactory factor recovery (*K* = 0.95, *FSC* = 0.95, Heywood cases = 0.05, *ES* = 4). Abbreviations: E = extraversion, O = openness, C = conscientiousness, N = neuroticism, A = agreeableness.

The subsampling study confirmed the findings of the simulation study with respect to the fact that a larger sample size is needed when extracting a larger number of factors. It should be noted that the subsampling study yielded moderately higher estimates of N compared to the simulations. For example, for (mean) $\lambda = 0.58$, $f = 3$, and $p = 27$, the subsampling study yielded an $N = 73$, whereas the respective value ($\lambda = 0.6$, $f = 3$, and $p = 24$) in the simulations (Table 2) was 55. This discrepancy can be attributed to the presence of model error as well as to the five-point Likert scale data: To obtain reliable correlations between variables, Likert scale data require a larger sample size than continuous normally distributed data.

4. Discussion and recommendations

The goal of this article is to offer a comprehensive overview of the conditions in which EFA can yield good quality results for small N. The simulations showed that, for the circumstances under which EFA is mostly applied (i.e., low to medium loadings, communalities, and a relatively large number of factors), a large sample size is required. However, when the data are well-conditioned (i.e., high λ, low f, high p), EFA can yield reliable solutions for sample sizes well below 50. In some conditions, sample sizes even smaller than 10 (beyond the smallest sample size of previous simulation studies) were sufficient. For example, when $\lambda = 0.8$, $f = 1$, $p = 24$, and the structure was simple, $N = 6$ was adequate. A small sample solution ($N = 17$, $\lambda = 0.8$, $f = 3$, $p = 24$) was markedly robust against single small distortions. Weakly determined factors and strong interfactor correlations negatively affected factor recovery, but even in the worst cases tested, factor recovery was still possible. The subsampling study confirmed the findings of the simulations with respect to the fact that a larger sample size is required when extracting a larger number of factors. For one extracted factor, a very small sample size (10–17) was adequate for satisfactory factor recovery.

An important issue when factor analysing small samples is whether it is possible to correctly estimate the number of factors. The simulations showed that when the structure is simple, in most conditions, small sample sizes can guarantee an adequate $ES > 4$. However, when deviating from a simple structure, researchers should be extra cautious when deciding on the number of factors, particularly if these factors are correlated.

This paper emphasizes that researchers should certainly not be encouraged to strive for small sample sizes. Large sample sizes are always beneficial and inevitably required when communalities are low. However, when factors are well-defined or their number is limited, small sample size EFA can yield reliable solutions. Thus, a small sample size should not be the sole criterion for rejecting EFA. Inversely, if one prefers, subjecting a small sample to EFA can be worthwhile and may possibly reveal valuable latent patterns. Considering that models are useful unless they are grossly wrong (MacCallum, 2003) and a small sample size factor analytic model is not per definition grossly wrong, applying factor analysis in an exploratory phase is better than rejecting EFA a priori. Obviously, the reliability and theoretical soundness of the solutions should be very carefully assessed.

4.1. Deviations from a simple structure

The present study investigated factor recovery when deviating from a simple structure, a situation most likely to occur in real data but which had not previously been systematically investigated. Past studies usually focused on one kind of distortion and on sample sizes larger than 50.

A number of studies (Boomsma & Hoogland, 2001; Gagné & Hancock, 2006; Gerbing & Anderson, 1987; Marsh et al., 1998) introduced an *ifc* of 0.3 in their simulation studies without, however, investigating the effect of different levels of *ifc*. Anderson and Gerbing (1984) examined two levels of *ifc*, but only with respect to improper solutions. It is surprising that the effect of *ifc* has not been exhaustively studied yet, considering that interfactor correlations are often present in psychological models. The current simulations investigated a range of *ifc* and revealed that a small *N* solution was able to sustain small interfactor correlations; factor recovery deteriorated considerably, however, when factors were strongly correlated.

Model error is usually considered as having no or only a small effect on factor recovery (MacCallum et al., 2001). The present results showed that when model error was small it was indeed of only little influence. A large model error, however, had a strong negative effect on factor recovery, particularly when its distribution was skewed in favour of the earlier factors in the sequence.

Gagné and Hancock (2006) found that when replacing higher with lower loadings (in an equal manner between factors), the number of improper solutions increased; appending loadings on the other hand, was beneficial. These findings are in agreement with the present simulations. Additionally, the present study showed that, when replacing high with low loadings, all four investigated indices deteriorated. When appending low loadings, on the other hand, the indices exhibited various tendencies, indicating the importance of assessing factor recovery by means of more than one index.

Past studies showed that unequal loadings within factors may cause an increased number of improper solutions (Anderson & Gerbing, 1984) and a less rapid improvement of factor recovery as *N* increases (Velicer & Fava, 1998). The current simulations showed that in the presence of unequal loadings within factors *K* and *ES* remained unaffected, whereas *FSC* increased up to near unity despite an increased likelihood of solutions exhibiting a Heywood case. The robustness of the small *N* solution to unequal loadings with respect to *K* and *ES* and the improvement of *FSC* are of interest because such conditions resemble real data.

Beauducel and Wittmann (2005) investigated factor recovery in the presence of secondary loadings and found that secondary loadings did not influence absolute indices of model fit but negatively affected incremental indices. The present study showed that factor recovery was robust against secondary loadings and that *ES* can even improve. This is an important result because small secondary loadings are inevitably present in real data and researchers consistently use the presence of secondary loadings as a reason to disregard items during the process of scale construction. It should be noted, however, that scale developers may still prefer simple structures to ensure that individual items do not end up in multiple subscales and subscale correlations do not risk becoming inflated.

Beauducel (2001) and Briggs and MacCallum (2003) investigated patterns with weak factors but both studies focused more on comparing the performance of various factor analytic methods rather than examining factor recovery. Ximénez (2006) investigated the recovery of weak factors using CFA and found that the recovery of weak factors may be troublesome if their loadings are small and the factors orthogonal. The present simulations investigated the effects of weak factors by means of unequal loadings between factors as well as by means of unequal *p*/*f*. When loadings were unequal, weak factors did not inhibit the recovery of the strong factors. For unequal *p*/*f*, the recovery of factors with low *p*/*f* deteriorated considerably. On the other hand, the recovery of the strong factors improved in terms of *K* (even if *p*/*f* of the strong factor was unchanged).

In summary, the present study investigated the effect of a wide range of single small distortions on EFA performance and showed that while a small *N* solution was relatively robust against distortions, factor recovery deteriorated strongly when factors were correlated or weak.

4.2. The effects of *p*, *f*, *p*/*f*, and *p*/*N*

The simulations showed that an increased *p* improved factor recovery, raising *K*, *FSC,* and *ES*, and reducing Heywood cases. An increased *p* was particularly beneficial for low λ patterns. A large number of references in the literature consider *p*/*f* as a strong factor analytic determinant. The present simulations confirm that *p*/*f* is an important criterion; lowering *p*/*f* had a negative influence on factor recovery. When *p*/*f* was equal between factors, however, *p* and *f* had clearly distinct effects on the quality of the factor solutions (see Figure 1); therefore the *p*/*f* ratio should not be considered as a comprehensive measure. For example, in a simple structure and for the same level of loadings (0.8), 2 factors and 12 variables (i.e., $p/f = 6$) required a minimum sample size of 11, whereas with 8 factors and 48 variables this minimum increased to 47. MacCallum et al. (2001) described a similar effect. They noticed that the effect of overdetermination on their empirical data was considerably weaker than in their Monte Carlo data. The difference was that in the empirical study the nature and number of factors were kept constant while the number of variables varied, whereas in the Monte Carlo study the number of variables was kept constant while the number of factors varied. The present study indicates that when *p*/*f* is equal, one should evaluate *p* and *f* separately instead of their ratio.

In some simulation conditions and in the subsampling study, the number of variables exceeded the sample size. Many factor analytic studies (e.g., Aleamoni, 1976), statistical packages, and factor analysis guidelines claim that the number of variables should never exceed the sample size. Contrary to this popular belief, Marsh and Hau (1999) reported no discontinuities in their simulation results when surpassing the $p = N$ barrier and suggested that there might be nothing special about such a barrier. The present simulations and the subsampling study concur with this view for all the investigated ranges of *p* and *N*. In fact, increasing the number of variables was beneficial, including when $p > N$. Moreover, in recent work, Robertson and Symons (2007) proved that $p > N$ is valid for maximum likelihood factor analysis. This method usually considers $p > N$ impossible because the covariance matrix turns nonpositive definite. Besides, as Bollen (2002) noted: "Resolution of this indeterminacy is theoretically possible under certain conditions …. (a) when the sample size (N) goes to infinity, (b) when the number of observed variables goes to infinity, and (c) when the squared multiple correlation for the latent variable goes to one and the predictors are observed variables" (p. 616). In other words, increasing the number of variables originates from the same striving for reducing factor indeterminacy as increasing the sample size. This is of importance for small sample sizes: when increasing *N* is not possible, one can attempt to append good quality variables, no matter if such a strategy may lead to a $p > N$ condition.

The simulations show that adding low loading variables considerably affected Tucker's congruence coefficient (*K*). This may then imply that only variables expected to load highly on a factor should be considered. Such a recommendation is only partially true as it could lead to the pitfall of "bloated specific" factors because highly loading variables can be also highly redundant (Boyle, 1991). Such variables lead to factors that are internally consistent but have low validity since they mask the presence or power of other factors and contaminate the entire factor structure.[5] In fact, the selected variables should be such that they assure validity, while being sufficiently diverse. The present simulations show that the factor score correlation coefficient (*FSC*) considerably improved when many variables were added, even when those variables had low factor loadings. In conclusion, we recommend in-

[5]) Cronbach's α was calculated for two conditions of the first simulation series (low loadings: $\lambda = 0.2$, $f = 2$, $p = 24$, $N = 1438$ and high loadings: $\lambda = 0.9$, $f = 2$, $p = 24$, $N = 6$). Although factor recovery was identical in those two conditions (see Table 2), average Cronbach's α amongst variables loading on the factor was 0.332 for the low loadings and 0.968 for the high loadings. This demonstrates that high internal consistency is not necessary for good factor recovery. A more detailed discussion of this issue can be found in Boyle (1991).

creasing the number of variables as much as possible, but only as long as this does not undermine the overall quality of the set.

4.3. Indices for assessing factor recovery

Indices used to evaluate the quality of factor solutions were *K*, *FSC*, Heywood cases, and *ES*. These indices exhibited varying tendencies and were sensitive to different determinants and distortions. The difference in the behaviour of *K* and *FSC* is attributed to their inherent nature. As an index of factor loadings similarity, *K* is influenced both by high and low loadings. *FSC* on the other hand, is an index of similarity of factor scores that are a weighted sum of the manifest variables. *FSC* monotonically increases with *p* because it benefits from added information in general. We conclude that *K* and *FSC* evaluate different aspects of factor recovery and recommend using them complementarily.

A number of studies have discussed the effects of sample size and *p*/*f* on the proportion of sample solutions exhibiting a Heywood case (or improper solutions) (e.g., Gerbing & Anderson, 1987; Marsh et al., 1998; Velicer & Fava, 1998). According to Velicer and Fava (1998), Heywood cases are more likely to occur when the sample size is small, *p*/*f* is limited, and the loadings are low. Boomsma and Hoogland (2001) noticed that high factor loadings can also lead to Heywood cases. In the present study, the Heywood cases occurred indeed when loadings were high. However, Heywood cases were not detrimental to factor recovery, as high *K* and high *FSC* could still be obtained in the unscreened solutions. This agrees with MacCallum et al. (1999, 2001), who carried out their simulations twice, once by screening out samples that yielded Heywood cases and again with unscreened data, and showed that there was virtually no difference in the results.

The present study not only included the Heywood cases but also used them as an index of factor recovery. Similarly, Briggs and MacCallum (2003) studied the behaviour of Heywood cases when comparing different methods of factor analysis. Gagné and Hancock (2006) used nonconvergent and Heywood cases as a primary index of model quality. Based on the results of the present study, we recommend using the proportion of solutions exhibiting Heywood cases as an additional index since it offers valuable information about the effect of determinants and distortions.

An important question when factor analysing small samples is whether the sample will consistently yield a correct decision as to the number of factors. Considering that none of the current methods for determining the number of factors is infallible (Fabrigar et al., 1999), *ES* was used to represent the size of the gap between the *f*-th and ($f+1$)-th eigenvalues. When making the simplifying assumption of normally distributed independent eigenvalues with equal standard deviations, an $ES = 4$ corresponds to a maximum of 95.5% correct classifications. To illustrate the eigenvalue gap size in real data, Figure 4 shows the scree plot of the subsampling study for

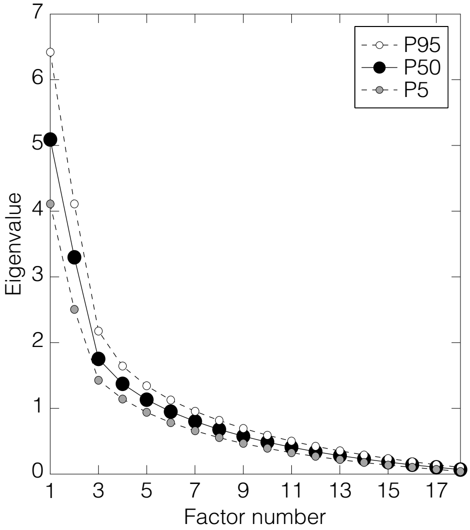


Figure 4. Scree plot of the subsampling study for $N = 39$ and $f = 2$ with the 5th, 50th, and 95th percentiles of the eigenvalues. The Cohen's *d* effect size (*ES*) between the second and third eigenvalue was 4.04. Applying the threshold at the optimal location (2.34) allowed for 96.5% correct estimations of $f = 2$. This figure is based on 100,000 repetitions.

$N = 39$ and $f = 2$. Here, between the second and third eigenvalue, *ES* was 4.04. Applying the threshold at the optimal location (2.34) allowed for 96.5% correct estimations of $f = 2$. A caveat is in order: *ES* does not identify the most appropriate number of factors, nor does it tell where the "large gap" or the "elbow" can be found in the scree plot. Rather, *ES* is a between-samples measure.

4.4. Deciding the number of factors

Deciding the "correct" number of factors has been the subject of many studies. As Bentler (2000) indicated: "Inevitably, due to the variety of possible criteria and methods of evaluating this question, in any empirical application there will never be precise unanimity among all researchers. This does not worry me too much because various models always fit in degrees …, and perhaps there may not even be a 'true' model" (p. 86). In other words, it is better to think in terms of "most appropriate" than "correct" number of factors. Yet, even when the common factor model holds exactly in the population and $ES > 4$ (such as in most current simulations), automatically estimating the correct number of factors is a challenge. We made several attempts to estimate the correct number of factors in the first series of simulations by using

Velicer's Minimum Average Partial (MAP; O'Connor, 2000), a Bayesian Information Criterion (BIC) (Hansen et al., 2001), an automatic scree test (Zhu & Ghodsi, 2006), and parallel analysis (O'Connor, 2000) (data not shown). Each of these methods was effective in many conditions, but none was successful in all conditions. A directly related topic is the effect of over- and underextraction (e.g., Fava & Velicer, 1992, 1996): Although it has been reported that the effect of overextraction can be stronger when N is small and λ low (Lawrence & Hancock, 1999), one may question whether factor misspecification is a small N problem per se or a matter of well- or ill-conditioned data. More research is needed on the strengths and weaknesses of procedures to determine the most appropriate number of factors.

4.5. Study limitations

The present study is not free of caveats or limitations. First, the simulation sample matrices were generated by a method described by Hong (1999), which produces normally distributed data and uses certain assumptions to generate model error (e.g., distribution of minor factors). Hong's method is a state-of-the-art procedure for introducing model error and interfactor correlations in a correlation matrix, based on the more commonly used Tucker-Koopman-Linn model (Tucker et al., 1969). Because of the normally distributed data, the simulations may have provided somewhat overoptimistic values for the minimum sample size compared to empirical data, as was also found in the subsampling study. Moreover, as Table 2 shows, the estimated minimum sample size would have been higher had factor recovery been assessed by using the 5th percentile of Tucker's congruence coefficient instead of its mean.

Second, although the simulated conditions corresponded to the ranges of determinants and distortions in psychological batteries, the conditions were of course far from exhaustive and might not be fully representative for all small sample conditions. For instance, one may conceive of structures that include combinations of distortions. Nonetheless, that factor recovery is possible in the presence of small distortions remains important for real applications.

Third, all correlation matrices were subjected to principal axis factoring and all loading matrices to oblique direct quartimin rotation. Different model fit procedures and rotations can have different effects on factor recovery. It is also possible that differently distorted matrices may have different favourable rotations. Those are issues that deserve further investigation.

Fourth, it should be noted that not just the factor recovery determines the quality of the factor analytic solution. As in any statistical analysis, the nature of both the sample and the variables involved remains among the most critical decisions (Fabrigar et al., 1999). A sample insufficiently representative of the population will distort the factor structure. Redundant variables can lead to bloated specific factors, obscuring the presence of more important factors. Irrelevant variables can also lead to spurious common factors (Fabrigar et al., 1999) and additional model error. Moreover, when the sample size is small, one should expect the standard error of loadings to be larger, which involves the risk of spurious high loadings.

Fifth, generalizing the findings to CFA should be done with great care. Although the communalities in CFA are usually higher due to variable selection (MacCallum et al., 2001), particular caution should be taken with respect to misspecification and stronger sources of model error. The presence of model error may alter the minimum required sample size for CFA. However, as MacCallum et al. (2001) noted, one can expect to find similar tendencies and determinants for CFA.

Finally, one may doubt that real data can satisfy the constraints of high communalities and loadings, or few factors. Moderate to weak communalities ranging between 0.4 and 0.7 (Costello & Osborne, 2005) or moderate to weak loadings ranging between 0.3 and 0.5 (Lingard & Rowlinson, 2006) are more common in behavioural or social data. The Big Five dataset of the subsampling study showed how indispensable a sufficiently large sample size is in such circumstances. However, cases involving high loadings do exist, for example, in neuroscience or psychosomatic research (e.g., Bailer et al., 2006; Gaines et al., 2006; Yuasa et al., 1995; with loadings up to 0.90 or 0.95). One-factor structures are not uncommon in scientific literature either, such as in psychometrics, psychiatry or epidemiology (e.g., general intelligence factor, self-concept, general distress factor, metabolic syndrome factor). Animal behaviour and behavioural genetics (Preacher & MacCallum, 2002) as well as evolutionary psychology (J.J. Lee, 2007) often offer data with high communalities and few factors. Outside the field of behavioural sciences, physics and chemistry can feature data with high reliability. Paradoxically, when high quality data are likely to occur, researchers seem to think there is no need to resort to latent structures and prefer deductive reasoning and mathematical modelling instead. The question is whether dismissing EFA in those cases is not accompanied by a weaker representation of the reality (Haig, 2005) when neglecting the latent pattern of the data. EFA is indeterminate by nature, but so is the empirical world.

Acknowledgements

We are grateful to Dr. Samuel D. Gosling for providing the dataset of the Big Five Inventory for the subsampling study.

References

Åberg, L., & Rimmö, P-A. (1998). Dimensions of aberrant driver behaviour. *Ergonomics, 41*, 39–56.

Acito, F., & Anderson, R.D. (1980). A Monté Carlo comparison of factor analytic methods. *Journal of Marketing Research, 17*, 228–236.

Advisory Group for Aerospace Research and Development. (1980). *Fidelity of simulation for pilot training* (Rep. No. AGARD-AR-159). Neuilly sur Seine, France: North Atlantic Treaty Organization.

Aleamoni, L.M. (1976). The relation of sample size to the number of variables in using factor analysis techniques. *Educational and Psychological Measurement, 36*, 879–883.

Alessi, S. (2004). Five keys to successful simulations. *Third International Sports Science Days*, 141–146. Retrieved 10 July 2008 from http://archiveouverte.campus-insep.net:81/archimede/INSEP/118/6-7-118-20060524-1.pdf

Allen, R.W., Guibert, M.R., Park, G.D., & Rosenthal, T.J. (2006). A user configurable PC platform for driver assessment and training. *Proceedings of the Driving Simulation Conference Asia/Pacific*, Tsukuba, Japan.

Allen, R.W., Marcotte, T.D., Rosenthal, T.J., & Aponso, B.L. (2005a). Driver assessment with measures of continuous control behavior. *Proceedings of the Third International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*, Rockport, ME, 165–172.

Allen, R.W., Park, G.D., Cook, M.L., & Fiorentino, D. (2007a). The effect of driving simulator fidelity on training effectiveness. *Proceedings of the Driving Simulation Conference North America*, Iowa City, IA.

Allen, R.W., Park, G.D., Cook, M.L., & Fiorentino, D. (2007b). A simulator for assessing older driver skills. *Advances in Transportation Studies; An International Journal, Special Issue*, 23–32.

Allen, R.W., Rosenthal, T.J., & Aponso, B.L. (2005b). Measurement of behavior and performance in driving simulation. *Proceedings of the Driving Simulation Conference North America*, Orlando, FL, 240–250.

Anderson, J.C., & Gerbing, D.W. (1984). The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika, 49*, 155–173.

Arnett, J.J. (1996). Sensation seeking, aggressiveness, and adolescent reckless behavior. *Personality and Individual Differences, 20*, 693–702.

Arnett, J.J. (2002). Developmental sources of crash risk in young drivers. *Injury Prevention, 8*, ii17–ii23.

Arrindell, W.A., & Van der Ende, J. (1985). An empirical test of the utility of the observations-to-variables ratio in factor and components analysis. *Applied Psychological Measurement, 9*, 165–178.

Backlund, P., Engström, H., Johannesson, M., & Lebram, M. (in press). Games for traffic education: An experimental study of a game-based driving simulator. *Simulation & Gaming*.

Bailer, J., Witthöft, M., & Rist, F. (2006). The chemical odor sensitivity scale: Reliability and validity of a screening instrument for idiopathic environmental intolerance. *Journal of Psychosomatic Research, 61*, 71–79.

Balogh, T., Forgacs, T., Agocs, T., Kovacs, P.T., Dobranyi, Z., Bouvier, E., & Balet, O. (2006). Concept of true 3D visualization in driving simulation with the HOLOVIZIO system. *Proceedings of the Driving Simulation Conference Europe*, Paris, France, 79–88.

Bauer, A., Dietz, K., Kolling, G., Hart, W., & Schiefer, U. (2001). The relevance of stereopsis for motorists: A pilot study. *Graefe's Archive for Clinical and Experimental Ophthalmology, 239*, 400–406.

Baughan, C.J. (2006). Review of the practical driving test. *Proceedings of the Novice Drivers Conference*. Retrieved 10 July 2008 from http://www.dft.gov.uk/pdf/pgr/roadsafety/drs/novicedrivers/conference/reviewofthepracticaldrivingtest

Baughan, C.J., Gregersen, N.P., Hendrix, M., & Keskinen, E. (2005). *Towards European standards for testing: Final Report*: Commission Internationale des Examens de Conduite Automobile CIECA.

Baughan, C.J., & Sexton, B. (2001). Driving tests: Reliability and the relationship between test errors and accidents. *Proceedings of the First International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*, Aspen, CO, 264–269.

Bearden, W.O., Sharma, S., & Teel, J.E. (1982). Sample size effects on chi square and other statistics used in evaluating causal models. *Journal of Marketing Research, 19*, 425–430.

Beauducel, A. (2001). On the generalizability of factors: The influence of changing contexts of variables on different methods of factor extraction. *Methods of Psychological Research Online, 6,* 69–96.

Beauducel, A., & Wittmann, W.W. (2005). Simulation study on fit indexes in CFA based on data with slightly distorted simple structure. *Structural Equation Modeling, 12*, 41–75.

Bentler, P.M. (2000). Rites, wrong, and gold in model testing. *Structural Equation Modeling, 7*, 82–91.

Bernaards, C.A., & Jennrich, R.I. (2005). Gradient projection algorithms and software for arbitrary rotation criteria in factor analysis. *Educational and Psychological Measurement, 65*, 676–696.

Bjørnskau, T., & Sagberg, F. (2005). What do novice drivers learn during the first months of driving? Improved handling skills or improved road user interaction? In G. Underwood (Ed.), *Traffic and Transport Psychology, Theory and Application* (pp.129–140). London: Elsevier.

Blaauw, G.J. (1982). Driving experience and task demands in simulator and instrumented car: A validation study. *Human Factors, 24*, 473–486.

Blana, E. (1996). *Driving simulator validation studies: A literature review* (Working Paper 480). University of Leeds, UK: Institute of Transport Studies.

Blockey, P.N., & Hartley, L.R. (1995). Aberrant driving behaviour: Errors and violations. *Ergonomics, 38*, 1759–1771.

Blomquist, G. (1986). A utility maximization model of driver traffic safety behavior. *Accident Analysis & Prevention, 18*, 371–375.

Boer, E.R. (1999). Car following from the driver's perspective. *Transportation Research Part F: Traffic Psychology and Behaviour, 2*, 201–206.

Boer, E.R. (2006). Driving simulator validation: Car following gap controllability. *Proceedings of the Driving Simulation Conference Asia/Pacific*, Tsukuba, Japan.

Boer, E.R., Kuge, N., & Yamamura, T. (2001). Affording realistic stopping behavior: A cardinal challenge for driving simulators. *Proceedings of the 1st Human-Centered Transportation Simulation Conference*, Iowa City, IA.

Boer, E.R., Ward, N.J., Manser, M.P., Yamamura, T., & Kuge, N. (2005). Driver performance assessment with a car following model. *Proceedings of the Third International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*, Rockport, ME, 433–440.

Boer, E.R., Yamamura, T., Kuge, N., & Girshick, A. (2000). Experiencing the same road twice: A driver centered comparison between simulation and reality. *Proceedings of Driving Simulation Conference*, Paris, France, 33–55.

Bollen, K.A. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology, 53*, 605–634.

Bonsall, P.W., & Palmer, I.A. (1997). Do time-based road-user charges induce risk-taking? – Results from a driving simulator. *Traffic Engineering & Control, 38*, 200–203, 208.

Boomsma, A. (1982). The robustness of LISREL against small sample sizes in factor analysis models. In K.G. Jöreskog & H. Wold (Eds.), *Systems under indirect observation: Causality, structure, prediction (part 1)* (pp. 149–173). Amsterdam: North-Holland.

Boomsma, A., & Hoogland, J.J. (2001). The robustness of LISREL modeling revisited. In R. Cudeck, S. du Toit, & D. Sörbom (Eds.), Structural equation modeling: Present and future. A festschrift in honor of Karl Jöreskog (pp. 139–168). Lincolnwood, IL: Scientific Software International.

Boschloo, H.W., Wieringa, P.A., Kuipers, J., De Winter, J.C.F., & Mulder, M. (2005). Driving behaviour in low-cost driving simulators: The influence of vibrations on braking and cornering behaviour. *Proceedings of the 24th European Annual Conference on Human Decision Making and Manual Control,* Athens, Greece.

Boydstun, L.E., Kessel, D.S., & Miller, J.M. (1980). Assessment of perceptually disabled individuals driving skills with a driving simulator. *Proceedings of the Human Factors Society 24th Annual Meeting*, Santa Monica, CA, 111–113.

Boyle, G.J. (1991). Does item homogeneity indicate internal consistency or item redundancy in psychometric scales? *Personality and Individual Differences, 12*, 291–294.

Brackstone, M., & McDonald, M. (1999). Car-following: A historical review. *Transportation Research Part F: Traffic Psychology and Behaviour, 2*, 181–196.

Breedveld, P., Stassen, H.G., Meijer, D.W., & Jakimowicz, J. (2000). Observation in laparoscopic surgery: Overview of impeding effects and supporting aids. *Journal of Laparoendoscopic & Advanced Surgical Techniques. Part A, 10*, 231–241.

Briggs, N.E., & MacCallum, R.C. (2003). Recovery of weak common factors by maximum likelihood and ordinary least squares estimation. *Multivariate Behavioral Research, 38*, 25–56.

Brookhuis, K.A. (2008). From ergonauts to infonauts: 50 years of ergonomics research. *Ergonomics, 51*, 55–58.

Brookhuis, K.A., & De Waard, D. (1993). The use of psychophysiology to assess driver status. *Ergonomics, 36*, 1099–1110.

Brookhuis, K.A., & De Waard, D. (2002). On the assessment of (mental) workload and other subjective qualifications. *Ergonomics, 45*, 1026–1030.

Brookhuis, K.A., De Waard, D., & Fairclough, S.H. (2003). Criteria for driver impairment. *Ergonomics, 46*, 433–445.

Brown, I.D. (1997). How traffic and transport systems can benefit from psychology. In T. Rothengatter & E. Carbonell Vaya (Eds.), *Traffic and Transport Psychology, Theory and Application* (pp. 9–19). Amsterdam: Pergamon.

Browne, M.W. (1968). A comparison of factor analytic techniques. *Psychometrika, 33*, 267–334.

Bürki-Cohen, J., Soja, N.N., & Longridge, T. (1998). Simulator platform motion – the need revisited. *The International Journal of Aviation Psychology, 8*, 293–317.

Cacciabue, P.C. (Ed.). (2007). *Modelling driver behaviour in automotive environments: Critical issues in driver interactions with intelligent transport systems*. London: Springer-Verlag.

Carsten, O. (2007). From driver models to modelling the driver: What do we really need to know about the driver? In P.C. Cacciabue (Ed.), *Modelling driver behaviour in automotive environments: Critical issues in driver interactions with intelligent transport systems* (pp. 105–120). London: Springer-Verlag.

Carstensen, G. (2002). The effect on accident risk of a change in driver education in Denmark. *Accident Analysis & Prevention, 34*, 111–121.

Cattell, R.B. (1978). *The scientific use of factor analysis in behavioral and life sciences.* New York: Plenum.

CBR (2003). *Jaarverslag 2003. Lijn brengen in verkeersveiligheid*, Rijswijk, the Netherlands. Retrieved 8 August 2008 from: www.cbr.nl/download/JaarverslagCBR2003.pdf

CBR (2005). *Review of 2005. On the land and water, and in the air*, Rijswijk, the Netherlands.

CBR (2007). *Result Form Driving Test B*. http://rijbewijs.cbr.nl/pdf/info%20USF%20kandidaat%20PPE_logo.pdf (Accessed 28 February 2008).

CBR (2008). *Pass rates of driving schools*. http://www.rijschoolgegevens.nl (Accessed April 2006/2007/2008).

CCV (2008). Examens beroepschauffeur: Publicaties [Tests professional drivers: Publications]. Retrieved 18 August 2008 from http://www.cbr.nl/10514.pp

Chalmé, S., Chaali, A., & Anceaux, F. (2006). Subjective risk and traffic violations: First results of an experimental study in a driving simulator. *Proceedings of the 25th European Annual Conference on Human Decision Making and Manual Control*, Valenciennes, France.

Chapman, P., Underwood, G., & Roberts, K. (2002). Visual search patterns in trained and untrained novice drivers. *Transportation Research Part F: Traffic Psychology and Behaviour, 5*, 157–167.

Chmarra, M.K., Klein, S., De Winter, J.C.F., Jansen, F-W., & Dankelman, J. (2008). *How to objectively classify residents based on their psychomotor laparoscopic skills?* Manuscript submitted for publication.

Christie, R. (2001). *The effectiveness of driver training as a road safety measure: A review of the literature* (Report No. 01/03). Melbourne, Australia: Royal Automobile Club of Victoria (RACV).

Clarke, D.D., Ward, P., & Truman, W. (2005). Voluntary risk taking and skill deficits in young driver accidents in the UK. *Accident Analysis & Prevention, 37*, 523–529.

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

Colom, R., Juan-Espinosa, M., Abad, F., Garca, L.F. (2000). Negligible sex differences in general intelligence. *Intelligence, 28*, 57–68.

Colombet, F., Dagdelen, M., Reymond, G., Pere, C., Merienne, F., & Kemeny, A. (2008). Motion cueing: What is the impact on the driver's behavior? *Proceedings of the Driving Simulation Conference Europe*, Monaco, 171–181.

Coluccia, E., & Louse, G. (2004). Gender differences in spatial orientation: A review. *Journal of Environmental Psychology, 24*, 329–340.

Comrey, A.L. (1973). *A first course in factor analysis*. New York: Academic Press.

Congdon, P. (1999). *VicRoads Hazard Perception Test: Can it predict accidents?* (Report CR-99-1). Camberwell, Australia: Australian Council for Educational Research.

Costello, A.B., & Osborne, J.W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation, 10*, 1–9.

Crinson, L.F., & Grayson, G.B. (2005). Profile of the British learner driver. In G. Underwood (Ed.), *Traffic and Transport Psychology, Theory and Application* (pp. 157–170). London: Elsevier.

Crossman, E.R.F.W. (1959). A theory of the acquisition of speed-skill. *Ergonomics, 2*, 153–166.

Cudeck, R., & MacCallum, M.C. (Eds.). (2007). *Factor analysis at 100: Historical developments and future directions*. Mahwah, NJ: Lawrence Erlbaum Associates.

Cunningham, D.W., Chatziastros, A., Von der Heyde, M., & Bülthoff, H.H. (2001). Driving in the future: Temporal visuomotor adaptation and generalization. *Journal of Vision, 1*, 88–98.

Cutting, J.E. (1997). How the eye measures reality and virtual reality. *Behavior Research Methods, Instruments, & Computers 29*, 27–36.

Cutting, J.E. (2000). Accuracy, scope, and flexibility of models. *Journal of Mathematical Psychology, 44*, 3–19.

Davis, B.T., & Green, P. (1995). *Benefits of sound for driving simulation: An experimental evaluation* (Report No. UMTRI-95–16). Ann Arbor, MI: The University of Michigan, Transportation Research Institute, Ann Arbor, MI.

Deery, H.A., & Fildes, B.N. (1999). Young novice driver subtypes: Relationship to high-risk behavior, traffic accident record, and simulator driving performance. *Human Factors, 41*, 628–643.

Deffenbacher, J.L., Deffenbacher, D.M., Lynch, R.S., & Richards, T.L. (2003). Anger, aggression, and risky behavior: A comparison of high and low anger drivers. *Behaviour Research and Therapy, 41*, 701–718.

De Groot, S., De Winter, J.C.F., Mulder, J.A., Kuipers, J., & Wieringa, P.A. (2006). The effects of route-instruction modality on driving performance in a simulator. *Proceedings of the 9th TRAIL congress*, Rotterdam, the Netherlands.

De Groot, S., De Winter, J.C.F., Mulder, M., & Wieringa, P.A. (2007). Didactics in simulator-based driver training: Current state of affairs and future potential. *Proceedings of the Driving Simulation Conference North America*, Iowa City, IA.

De Groot, S., De Winter, J.C.F., & Wieringa, P.A. (2008). *Motion-cueing in a fixed-base driving simulator: 11 low-cost systems evaluated*. Manuscript in preparation.

De Pelsmacker, P., & Janssens, W. (2005). The effect of norms, attitudes and habits on speeding behavior: Scale development and model building and estimation. *Accident Analysis & Prevention, 39*, 6–15.

De Winter, J.C.F. (2008). *Towards a model of driver behaviour*. Unpublished manuscript. (chapter 2 of this thesis)

De Winter, J.C.F., De Groot, S., Dankelman, J., Wieringa, P.A., Van Paassen, M.M., & Mulder, M. (2008a). Advancing simulation-based driver training: Lessons learned and future perspectives. *Proceedings of the 10th International Conference on Human-Computer Interaction with Mobile Devices and Services (Mobile HCI 2008)*, Amsterdam, the Netherlands, 459–464. (chapter 5 of this thesis)

De Winter, J.C.F., De Groot, S., Mulder, M., & Wieringa, P.A. (2007a). The fun of engineering: A motion seat in a driving simulator. *Proceedings of the Driving Simulation Conference North America*, Iowa City, IA. (chapter 7 of this thesis)

De Winter, J.C.F., De Groot, S., Mulder, M., & Wieringa P.A. (2008b). *Issues in on-road driver training and prospects for simulation-based training*. Manuscript submitted for publication. (chapter 1 of this thesis)

De Winter, J.C.F., De Groot, S., Mulder, M., Wieringa, P.A., & Dankelman, J. (2008c). The search for higher fidelity in fixed-base driving simulation: Six feedback systems evaluated. *Proceedings of the Driving Simulation Conference Europe*, Monaco, 183–192. (chapter 8 of this thesis)

De Winter, J.C.F., De Groot, S., Mulder, M., Wieringa, P.A., Dankelman, J., & Mulder, J.A. (in press-a). Relationships between driving simulator performance and driving test results. *Ergonomics*. (chapter 4 of this thesis)

De Winter, J.C.F., De Groot, S., Van Loenhout, M.J., Van Leeuwen, A., Do, P., Wieringa, P.A., & Mulder, M. (2008d). Feedback on mirror-checking during simulation-based driver training. *Proceedings of the 27th European Annual Conference on Human Decision Making and Manual Control*, Delft, the Netherlands. (chapter 9 of this thesis)

De Winter, J.C.F., Dodou, D., De Groot, S., & Wieringa, P.A. (2008e). *Manual control theory: Learning by experiencing.* Manuscript submitted for publication.

De Winter, J.C.F., Dodou, D., & Wieringa, P.A. (in press-b). Exploratory factor analysis with small sample sizes. *Multivariate Behavioral Research*. (appendix B of this thesis)

De Winter, J.C.F., Houtenbos, M., Wieringa, P.A., Mulder, J.A., Kuipers, J., & De Groot, S. (2006a). Individual characteristics affecting intersection behavior in a driving simulator. *Proceedings of the 25th European Annual Conference on Human Decision Making and Manual Control*, Valenciennes, France.

De Winter, J.C.F., Kuipers, J., Venekamp, D.W., Wieringa, P.A., Mulder, M., & Van Paassen, M.M. (2006b). Performance assessment during simulation-based driver training. *Proceedings of the 16th World Congress of the International Ergonomics Association*, Maastricht, the Netherlands.

De Winter, J.C.F., Mulder, M., De Groot, S., & Wieringa, P.A. (2006c). Factor analysis for driving assessment applied to car following in a simulator. *Proceedings of the 9th TRAIL congress*, Rotterdam, the Netherlands.

De Winter, J.C.F., Mulder, M., Van Paassen, M.M., Abbink, D.A., & Wieringa, P.A. (2008f). A two-dimensional weighting function for a driver-assistance system. *IEEE Transactions on Systems, Man, and Cybernetics – Part B: Cybernetics, 38*, 189–195.

De Winter, J.C.F., & Wieringa, P.A. (2008). Gender differences in driver's license theory test scores in the Netherlands. *Journal of Safety Research*, *39*, 413–416. (appendix A of this thesis)

De Winter, J.C.F., Wieringa, P.A., Dankelman, J., Mulder, M., Van Paassen, M.M., & De Groot, S. (2007b). Driving simulator fidelity and training effectiveness. *Proceedings of the 26th European Annual Conference on Human Decision Making and Manual Control*, Lyngby, Denmark. (chapter 6 of this thesis)

De Winter, J.C.F., Wieringa, P.A., Kuipers, J., Mulder, J.A., & Mulder, M. (2007c). Violations and errors during simulation-based driver training. *Ergonomics, 50*, 138–158. (chapter 3 of this thesis)

De Winter, J.C.F., Wieringa, P.A., Kuipers, J., Mulder, M., Van Paassen, M.M., & Dankelman, J. (2006d). Driver assessment during self and forced-paced simulation-based training. *Proceedings of Driving Simulation Conference Europe*, Paris, France, 157–166.

Diete, F. (2008). *Evaluation of a simulator based, novice driver risk awareness training program*. Master's thesis, University of Massachussetts Amherst.

Dols, J.F., Pardo, J., Falkmer, T., Uneken, E., & Verwey, W. (2001). The Trainer Project: A new simulator-based driver training curriculum. *Proceedings of the First International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*, Aspen, CO, 119–124.

Drascic, D. (1991). Skill acquisition and task performance in teleoperation using monoscopic and stereoscopic video remote viewing. *Proceedings of the Human Factors Society 35th Annual Meeting*, San Francisco, CA, 1367–1371.

Drews, F.A., Strayer, D.L., Uchino, B.N., & Smith, T.W. (2003). On the fast lane to road rage. *Proceedings of the Second International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design,* Park City, UT, 194–198.

Drummond, A.E. (1989). *An overview of novice driver performance issues. A literature review* (Report No. 9). Clayton, Australia: Monash University, Accident Research Centre.

Duncan, J., Williams, P., & Brown, I. (1991). Components of driving skill: Experience does not mean expertise. *Ergonomics, 34*, 919–937.

Eichenlaub, J.B. (2007). Passive method of eliminating accommodation/convergence disparity in stereoscopic head mounted displays. *Proceedings of the SPIE-IS&T Electronic Imaging*, 517–529.

Elander, J., West, R., & French, D. (1993). Behavioral correlates of individual differences in road-traffic crash risk: An examination of methods and findings. *Psychological Bulletin, 113*, 279–294.

Elvik, R., & Vaa, T. (2004). *Handbook of road safety measures*. Oxford: Elsevier Science.

Emmerson, K. (2008). *Learning to drive: The evidence* (No. 87). London, UK: Department for Transport.

Ernst, M.O., & Bülthoff, H.H. (2004). Merging the senses into a robust percept. *Trends in Cognitive Sciences, 8*, 162–169.

European Commission (2007). *Summary and publication of best practices in road safety in the member states (SUPREME). Part F2. Thematic report: Driver education, Training & Licensing.* Retrieved 10 July 2008 from http://ec.europa.eu/transport/roadsafety_library/publications/supreme_f2_thematic_report_driver_education_training_licensing.pdf

European Transport Safety Council (2003). *Assessing risk and setting targets in transport safety programmes*. Brussels, Belgium.

Evans, L. (1970). Speed estimation from a moving automobile. *Ergonomics, 13*, 219–230.

Evans, L. (2004). *Traffic safety*. Bloomfield Hills, MI: Science Serving Society.

Evans, L. (2006). Innate sex differences supported by untypical traffic fatalities. *Chance, 19*, 10–15.

Everett, J.E. (1983). Factor comparability as a means of determining the number of factors and their rotation. *Multivariate Behavioral Research, 18*, 197–218.

Fabrigar, L.R., Wegener, D.T., MacCallum, R.C., & Strahan, E.J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*, 272–299.

Fava, J.L., & Velicer, W.F. (1992). The effects of overextraction on factor and component analysis. *Multivariate Behavioral Research, 27*, 387–415.

Fava, J.L., & Velicer, W.F. (1996). The effects of underextraction in factor and component analysis. *Educational and Psychological Measurement, 56*, 907–929.

Fikkert, W., Heylen, D., Van Dijk, B., Nijholt, A., Kuipers, J., & Brugman, A. (2006). Estimating the gaze point of a student in a driving simulator. *Proceedings of the 6th*

International Conference on Advanced Learning Technologies, 497–501.

Fisher, D.L., Pollatsek, A.P., & Pradhan, A. (2006). Can novice drivers be trained to scan for information that will reduce their likelihood of a crash? *Injury Prevention, 12*, i25–i29.

Fitts, P.M. (1954). The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology*, *47*, 381–391.

Flach, J.M. (1990). Control with an eye for perception: Precursors to an active psychophysics. *Ecological Psychology, 2*, 83–111.

Flach, J.M., Dekker, S., & Stappers, P.J. (2008). Playing twenty questions with nature (the surprise version): Reflections on the dynamics of experience. *Theoretical Issues in Ergonomics Science, 9*, 125–154.

Flipo, A. (2000). TRUST: The truck simulator for training. *Proceedings of the Driving Simulation Conference*, Paris, France, 293–302.

Forsyth, E. (1992). *Cohort study of learner and novice drivers. Part 1: Learning to drive and performance in the driving test* (Report No. 338). Crowthorne, UK: Transport Research Laboratory.

French, D.J., West, R.J., Elander, J., & Wilding, J.M. (1993). Decision-making style, driving style, and self-reported involvement in road traffic accidents. *Ergonomics, 36*, 627–644.

Frex Japan Trading (2008). FrexGP SimConMOTION. http://www.frex.com/gp (Accessed 10 July 2008).

Fuller, R. (2005). Towards a general theory of driver behaviour. *Accident Analysis & Prevention, 37*, 461–472.

Gagné, P., & Hancock, G.R. (2006). Measurement model quality, sample size, and solution propriety in confirmation factor models. *Multivariate Behavioral Research, 41*, 65–83.

Gaines, J.J., Shapiro, A., Alt, M., & Benedict, R.H.B. (2006). Semantic clustering indexes for the Hopkins Verbal Learning Test-Revised: Initial exploration in elder control and dementia groups. *Applied Neuropsychology, 13*, 213–222.

Gawron, V.J., & Ranney, T.A. (1988). The effects of alcohol dosing on driving performance on closed course and in a driving simulator. *Ergonomics, 31*, 1219–1244.

Gebers, M.A., & Peck, R.C. (2003). Using traffic conviction correlates to identify high accident-risk drivers. *Accident Analysis & Prevention, 35*, 903–912.

George, C.F.P. (2003). Driving simulators in clinical practice. *Sleep Medicine Reviews, 7*, 311–320.

Gerbing, D.W., & Anderson, J.C. (1987). Improper solutions in the analysis of covariance structures: Their interpretability and a comparison of alternate respecifications. *Psychometrika, 52*, 99–111.

Geweke, J.F., & Singleton, K.J. (1980). Interpreting the likelihood ratio statistic in factor models when sample size is small. *Journal of the American Statistical Association, 75*, 133–137.

Gibson, J.J., & Crooks, L.E. (1938). A theoretical field-analysis of automobile-driving. *The American Journal of Psychology, 51*, 453–471.

Ginzburg, L.R., & Jensen, C.X.J. (2004). Rules of thumb for judging ecological theories. *Trends in Ecology and Evolution, 19*, 121–126.

Glendon, A.I. (2007). Driving violations observed: An Australian study. *Ergonomics, 50*, 1159–1182.

Golob, T.F. (2003). Structural equation modeling for travel behavior research. *Transportation Research Part B: Methodological, 37*, 1–25.

Gorsuch, R.L. (1974). *Factor analysis*. Philadelphia: Saunders.

Gosling, S.D., Vazire, S., Srivastava, S., & John, O.P. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. *American Psychologist, 59*, 93–104.

Gottfredson, L.S. (1997). Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography. *Intelligence, 24*, 13–23.

Green, P. (2005). How driving simulator data quality can be improved. *Proceedings of the Driving Simulation Conference North America*, Orlando, FL, 210–220.

Green Dino Virtual Realities (2007). Dutch Driving Simulator. http://www.dutchsimulators.com (Accessed 10 July 2008).

Green Dino Virtual Realities (2008). Dutch Driving Simulator. http://www.drivemasters.nl (Accessed 10 July 2008).

Greenberg, J., Artz, B., & Cathey, L. (2003). The effect of lateral motion cues during simulated driving. *Proceedings of the Driving Simulation Conference North America*, Dearborn, MI.

Gregersen, N.P., Berg, H-Y., Engström, I., Nolén, S., Nyberg, A., & Rimmö, P-A. (2000). Sixteen years age limit for learner drivers in Sweden – an evaluation of safety effects. *Accident Analysis & Prevention, 32*, 25–35.

Gregersen, N.P., Nyberg, A., & Berg, H-Y. (2003). Accident involvement among learner drivers – an analysis of the consequences of supervised practice. *Accident Analysis & Prevention, 35*, 725–730.

Grayson, G.B., Maycock, G., Groeger, J.A., Hammond, S.M., & Field, D.T. (2003). *Risk, hazard perception and perceived control* (TRL Report 560). Crowthorne, UK: Transport Research Laboratory.

Groeger, J.A. (2000a). *Understanding driving: Applying cognitive psychology to a complex everyday task*. Philadelphia: Psychology Press.

Groeger, J.A. (2000b). Fast learners: Once a speeder, always a speeder? *Proceedings of the 10th seminar on behavioural research in road safety*, Esher, Surrey, 144–151.

Groeger, J.A. (2001). Learning to drive with parents and professionals: A conundrum resolved? *Proceedings of the Novice Drivers Conference*, Bristol, UK.

Groeger, J.A. (2006). Youthfulness, inexperience, and sleep loss: The problems young drivers face and those they pose for us. *Injury Prevention, 12*, i19–i24.

Groeger, J.A., & Banks, A.P. (2007). Anticipating the content and circumstances of skill transfer: Unrealistic expectations of driver training and graduated licensing? *Ergonomics, 50*, 1250–1263.

Groeger, J.A., & Clegg, B.A. (2007). Systematic changes in the rate of instruction during driver training. *Applied Cognitive Psychology, 21*, 1229–1244.

Groeger, J.A., & Rothengatter, T. (1998). Traffic psychology and behaviour. *Transportation Research Part F: Traffic Psychology and Behaviour, 1*, 1–9.

Grünwald, P. (2000). Model selection based on minimum description length. *Journal of Mathematical Psychology*, *44*, 133–152.

Guilford, J.P. (1954). *Psychometric methods* (2nd ed.). New York: McGraw-Hill.

Gulian, E., Matthews, G., Glendon, A.I., Davies, D.R., Debney, L.M. (1989). Dimensions of driver stress. *Ergonomics, 32*, 585–602.

Guo, K., & Guan, H. (1993). Modelling of driver/vehicle directional control system. *Vehicle System Dynamics, 22*, 141–184.

Haig, B.D. (2005). Exploratory factor analysis, theory generation, and scientific method. *Multivariate Behavioral Research, 40*, 303–329.

Hair, J.F., Anderson, R.E., Tatham, R.L., & Grablowsky, B.J. (1979). *Multivariate data analysis.* Tulsa, OK: Pipe Books.

Hale, A.R., Stoop, J., & Hommels, J. (1990). Human error models as predictors of accident scenarios for designers in road transport systems. *Ergonomics, 33*, 1377–1387.

Hall, J., & West., R. (1996). Role of formal instruction and informal practice in learning to drive. *Ergonomics, 39*, 693–706.

Hancock, P.A. (1999). Is car following the real question – are equations the answer? *Transportation Research Part F: Traffic Psychology and Behaviour, 2*, 197–199.

Hansen, L.K., Larsen, J., & Kolenda, T. (2001). Blind detection of independent dynamic components. *IEEE International Conference on Acoustics, Speech, and Signal Processing, 5*, 3197–3200.

Harrison, W.A. (1999). The limited potential of training for learner drivers: A view from the psychologists' lab. *Paper presented to the Australian College of Road Safety Young Driver Conference.*

Hartos, J.L., Eitel, P., Haynie, D.L., & Simons-Morton, B.G. (2000). Can I take the car?: Relations among parenting practices and adolescent problem-driving practices. *Journal of Adolescent Research, 15*, 352–367.

Hatakka, M., Keskinen, E., Baughan, C., Goldenbeld, C., Gregersen, N.P., Groot, H., et al. (Eds.). (2003). *Basic driver training: New models.* Finland: University of Turku.

Hatakka, M., Keskinen, E., Gregersen, N.P., Glad, A., & Hernetkoski, K. (2002). From control of the vehicle to personal self-control; broadening the perspectives to driver education. *Transportation Research Part F: Traffic Psychology and Behaviour, 5*, 201–215.

Hatakka, M., Keskinen, E., Katila, A., & Laapotti, S. (1997). Self-reported driving habits are valid predictors of violations and accidents. In T. Rothengatter & E. Carbonell Vaya (Eds.), *Traffic and Transport Psychology, Theory and Application* (pp. 295–303). Amsterdam: Pergamon.

Hauber, A.R. (1980). The social psychology of driving behaviour and the traffic environment: Research on aggressive behaviour in traffic. *International Review of Applied Psychology, 29*, 461–474.

Hays, R.T., Jacobs, J.W., Prince, C., & Salas, E. (1992). Flight simulator training effectiveness: A meta-analysis. *Military Psychology, 4*, 63–74.

Hazevoet, A., & Vissers, J.A.M.M. (2005). *Periodiek Rijopleidingsonderzoek 2004–2005; Algemene vraagstelling*. TT 04-078. Traffic Test, Veenendaal.

Henson, R.K., & Roberts, J.K. (2006). Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educational and Psychological Measurement, 66*, 393–416.

Hoedemaeker, M., & Brookhuis, K.A. (1998). Behavioural adaptation to driving with an adaptive cruise control (ACC). *Transportation Research Part F: Traffic Psychology and Behaviour, 1*, 95–106.

Hoeschen, A., Verwey, W., Bekiaris, E., Knoll, C., Widlroither, H., De Waard, D., et al. (2001). *TRAINER: Inventory of driver training needs and major gaps in the relevant training procedures* (Deliverable No 2.1). Brussels, Belgium: European Commission.

Hollnagel, E., Nåbo, A., & Lau, I.V. (2003). A systemic model for driver-in-control. *Proceedings of the Second International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design,* Park City, UT, 86–91.

Hong, S. (1999). Generating correlation matrices with model error for simulation studies in factor analysis: A combination of the Tucker-Koopman-Linn model and Wijsman's algorithm. *Behavior Research Methods, Instruments, & Computers, 31*, 727–730.

Horvath, P., & Zuckerman, M. (1993). Sensation seeking, risk appraisal, and risky behavior. *Personality and Individual Differences, 14*, 41–52.

Houtenbos, M. (2008). *Expecting the unexpected: A study of interactive driving behaviour at intersections*. Doctoral dissertation, Delft University of Technology, the Netherlands.

Huguenin, R.D. (1988). The concept of risk and behaviour models in traffic psychology. *Ergonomics, 31*, 557–569.

Huguenin, R.D., & Rumar, K. (2001). Models in Traffic Psychology. In P-E. Barjonet (Ed.), *Traffic psychology today* (pp. 31–59). Boston: Kluwer Academic Publishers.

Iversen, H., & Rundmo, T. (2002). Personality, risky driving and accident involvement among Norwegian drivers. *Personality and Individual Differences, 33*, 1251–1263.

Iversen, H., & Rundmo, T. (2004). Attitudes towards traffic safety, driving behaviour and accident involvement among the Norwegian public. *Ergonomics, 47*, 555–572.

Jacobs, A.M., & Grainger, J. (1994). Models of visual word recognition: Sampling the state of the art. *Journal of Experimental Psychology: Human Perception and Performance, 20*, 1311–1334.

Jackson, D.L. (2001). Sample size and number of parameter estimates in maximum likelihood confirmatory factor analysis: A Monte Carlo investigation. *Structural Equation Modeling, 8*, 205–223.

Jagacinski, R.J., & Flach, J.M. (2003). *Control theory for humans: Quantitative approaches to modeling performance*. Mahwah, NJ: Lawrence Erlbaum Associates.

Jamson, H.A., Horrobin, A.J., & Auckland, R.A. (2007). Whatever happened to the LADS? Design and development of the new University of Leeds driving simulator. *Proceedings of the Driving Simulation Conference North America*, Iowa City, IA.

Janssen, W.H. (1994). Seat-belt wearing and driving behavior: An instrumented-vehicle study. *Accident Analysis & Prevention, 26*, 249–261.

Janssen, W.H., & Tenkink, E. (1988). Considerations on speed selection and risk homeostasis in driving. *Accident Analysis & Prevention, 20*, 137–142.

Jennrich, R.I., & Trendafilov, N.T. (2005). Independent component analysis as a rotation method: A very different solution to Thurstone's box problem. *British Journal of Mathematical and Statistical Psychology, 58*, 199–208.

Jensen, A.R. (2006). *Clocking the mind: Mental chronometry and individual differences*. Amsterdam: Elsevier.

Johnson, D.M. (2005). *Introduction to and review of simulator sickness research* (Research Report 1832). Fort Rucker, AL: U.S. Army Research Institute.

Johnson, D.M., & Stewart, J.E., II (1999). Use of virtual environments for the acquisition of spatial knowledge: Comparison among different visual displays. *Military Psychology, 11*, 129–148.

Johnston, I.R. (1992). Traffic safety education: Panacea, prophylactic or placebo. *World Journal of Surgery, 16*, 374–378.

Jones, M.H. (1973). *California Driver Training Evaluation Study. Final report to the California Legislature.* National Highway Traffic Safety Administration, Department of Transportation, Washington DC.

Kano, Y., Abe, Y., & Abe, M. (2005). Steering system parameter adaptation to handling characteristics of human driver. *Proceedings of the 24th European Annual Conference on Human Decision Making and Manual Control*, Athens, Greece.

Kaplunovsky, A.S. (2005). Factor analysis in environmental studies. *HAIT Journal of Science and Engineering B, 2*, 54–94.

Kappé, B., & Van Emmerik, M.L. (2005). *The use of driving simulators for initial driver training and testing* (Report TNO-DV3 2005 C114). Soesterberg: the Netherlands, TNO Defence, Security and Safety.

Kappé, B., Van Emmerik, M.L., Van Winsum, W., & Rozendom, A. (2003). Virtual instruction in driving simulators. *Proceedings of the Driving Simulation Conference North America*, Dearborn, MI.

Käppler, W.D. (1993). Views on the role of simulation in driver training. *Proceedings of the 12th European Annual Conference on Human Decision Making and Manual Control*, Kassel, Germany, *5*, 12–17.

Kaptein, N.A., Theeuwes, J., & Van der Horst, R. (1996). Driving simulator validity: Some considerations. *Transportation Research Record, 1550*, 30–36.

Katila, A., Keskinen, E., & Hatakka, M. (1996). Conflicting goals of skid training. *Accident Analysis & Prevention, 28*, 785–789.

Kemeny, A. (2000). Simulation and perception of movement. *Proceedings of the Driving Simulation Conference*, Paris, France, 13–22.

Kemeny, A., & Panerai, F. (2003). Evaluating perception in driving simulation experiments. *Trends in Cognitive Sciences, 7*, 31–37.

Kennedy R.S., Lane, N.E., Berbaum, K.S., & Lilienthal, M.G. (1993). Simulator Sickness Questionnaire: An enhanced method for quantifying simulator sickness. *The International Journal of Aviation Psychology, 3*, 203–220.

Ker, K., Roberts, I., Collier, T., Beyer, F., Bunn, F., & Frost, C. (2005). Post-licence driver education for the prevention of road traffic crashes: A systematic review of randomised controlled trials. *Accident Analysis & Prevention, 37*, 305–313.

Kim, J., Matsui, Y., Hayakawa, S., Suzuki, T., Okuma, S., & Tsuchida, N. (2005). Acquisition and modeling of driving skills by using three dimensional driving simulator. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, E88-A*, 770–778.

Kim, K.E. (1996). Differences between male and female involvement in motor vehicle collisions in Hawaii, 1986–1993. *Women's Travel Issues: Proceedings from the Second National Conference*, Baltimore, MD, 518–528.

Kleinman, D.L., Baron, S., & Levison, W.H. (1970). An optimal control model of human response part I: Theory and validation. *Automatica, 6*, 357–369.

Knapp, T.R., & Sawilowsky, S.S. (2005). Letter to the editor. *Journal of Nursing Measurement, 12*, 95–96.

Kontogiannis, T., Kossiavelou, Z., & Marmaras, N. (2002). Self-reports of aberrant behaviour on the roads: Errors and violations in a sample of Greek drivers. *Accident Analysis & Prevention, 34*, 381–399.

Koonce, J.M. (1979). Predictive validity of flight simulators as a function of simulator motion. *Human Factors, 21*, 215–223.

Kouwenberg, S. (2005). *Project Driving Force: Add a challenge to driving to retain attention and prevent recklessness* (Project Report for client Green Dino Virtual Realities). Eindhoven University of Technology, the Netherlands.

Kuiken, M.J., Miltenburg, P.G.M., & Van Winsum, W. (1992). Drivers' reactions to an intelligent driver support system (GIDS) implemented in a driving simulator. *Proceedings of the Third International Conference on Vehicle Navigation and Information Systems*, Rockport, ME, 176–181.

Laapotti, S., & Keskinen, E. (2004). Has the difference in accident patterns between male and female drivers changed between 1984 and 2000? *Accident Analysis & Prevention, 36*, 577–584.

Lajunen, T., Parker, D., & Stradling, S.G. (1998). Dimensions of driver anger, aggressive and highway code violations and their mediation by safety orientation in UK drivers. *Transportation Research Part F: Traffic Psychology and Behaviour, 1*, 107–121.

Lajunen, T., Parker, D., & Summala, H. (2004). The Manchester Driver Behaviour Questionnaire: A cross-cultural study. *Accident Analysis & Prevention*, *36*, 231–238.

Lajunen, T., & Summala, H. (1995). Driving experience, personality, and skill and safety-motive dimensions in drivers' self-assessments. *Personality and Individual Differences, 19*, 307–318.

Lambooij, M., IJsselsteijn, W., & Heynderickx, I. (2007). Stereoscopic displays and visual comfort: A review. *SPIE Newsroom*. Retrieved 10 July 2008 from http://spie.org/documents/Newsroom/Imported/0648/0648-2007-03-08.pdf

Lawrence, F.R., & Hancock, G.R. (1999). Conditions affecting integrity of a factor solution under varying degrees of overextraction. *Educational and Psychological Measurement, 59*, 549–579.

Lee, H.C. (2003). The validity of driving simulator to measure on-road driving performance of older drivers. *Transport Engineering in Australia, 8*, 89–100.

Lee, J.D. (2004). Simulator fidelity: How low can you go? *Proceedings of the 48th Annual Meeting of the Human Factors and Ergonomics Society*, Santa Monica, CA.

Lee, J.D. (2007). Technology and teen drivers. *Journal of Safety Research, 38*, 203–213.

Lee, J.J. (2007). A *g* beyond Homo sapiens? Some hints and suggestions. *Intelligence, 35*, 253–265.

Levine, M.S. (1977). *Canonical analysis and factor comparison*. Thousand Oaks, CA: Sage Publications.

Lew, H.L., Poole, J.H., Lee, E.H., Jaffe, D.L., Huang, H-C., & Brodd, E. (2005). Predictive validity of driving-simulator assessments following traumatic brain injury: A preliminary study. *Brain Injury, 19*, 177–188.

Lin, C-T., Wu, R-C., Jung, T-P., Liang, S-F., & Huang, T-Y. (2005). Estimating driving performance based on EEG spectrum analysis. *EURASIP Journal on Applied Signal Processing, 19*, 3165–3174.

Lingard, H., & Rowlinson, S. (2006). Letter to the editor. *Construction Management and Economics, 24*, 1107–1109.

Lorenz, B., & Manzey, D. (2001). Geschlechtsunterschiede bei Wahlreaktionsleistungen im eignungsdiagnostischen Kontext [Gender differences in choice reaction time performance in the context of aptitude testing]. *Zeitschrift für Differentielle und Diagnostische Psychologie, 22*, 173–184.

Lorenzo-Seva, U., & Ten Berge, J.M.F. (2006). Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology, 2*, 57–64.

Luke, T., Parkes, A.M., & Walker, R. (2006). The effect of visual properties of the simulated environment on simulator sickness and driver behaviour. *Proceedings of the Driving Simulation Conference Europe, Paris, France*, 253–262.

Lund, A.K., & Williams, A.F. (1985). A review of the literature evaluating the defensive driving course. *Accident Analysis & Prevention, 17*, 449–460.

Lund, A.K., Williams, A.F., & Zador, P. (1986). High school driver education: Further evaluation of the DeKalb County study. *Accident Analysis & Prevention, 18*, 349–357.

Lundqvist, A., Gerdle, B., & Rönnberg, J. (2000). Neuropsychological aspects of driving after a stroke – in the simulator and on the road. *Applied Cognitive Psychology, 14*, 135–150.

Luursema, J.-M., Verwey, W.B., Kommers, P.A.M., & Annema, J-H. (2008). The role of stereopsis in virtual anatomical learning. *Interacting with Computers*, *20*, 455–460.

MacAdam, C.C. (2003). Understanding and modeling the human driver. *Vehicle System Dynamics, 40*, 101–134.

MacCallum, R.C. (2003). 2001 Presidential Address. Working with imperfect models. *Multivariate Behavioral Research, 38*, 113–139.

MacCallum, R.C., & Tucker, L.R. (1991). Representing sources of error in the common factor model: Implications for theory and practice. *Psychological Bulletin, 109*, 502–511.

MacCallum, R.C., Widaman, K.F., Preacher, K.J., & Hong, S. (2001). Sample size in factor analysis: The role of model error. *Multivariate Behavioral Research, 36*, 611–637.

MacCallum, R.C., Widaman, K.F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4*, 84–99.

Males, M. (2007). California's graduated driver license law: Effect on teenage drivers' deaths through 2005. *Journal of Safety Research, 38*, 651–659.

Marsh, H.W., & Hau, K.T. (1999). Confirmatory factor analysis: Strategies for small sample sizes. In Hoyle, R.H. (Ed.), *Statistical issues for small sample research* (pp. 251–284). Thousand Oaks, CA: Sage.

Marsh, H.W., Hau, K.T., Balla, J.R., & Grayson, D. (1998). Is more ever too much: The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research, 33*, 181–220.

Masten, S.V., & Peck, R.C. (2004). Problem driver remediation: A meta-analysis of the driver improvement literature. *Journal of Safety Research, 35*, 403–425.

Matthews, G., & Desmond, P.A. (1997). Personality and multiple dimensions of task-induced fatigue: A study of simulated driving. *Personality and Individual Differences, 25*, 443–458.

Matthews, G., Dorn, L., Hoyes, T.W., Davies, D.R., Glendon, A.I., Taylor, R.G. (1998). Driver stress and performance on a driving simulator. *Human Factors, 40*, 136–149.

Maycock, G., & Forsyth, E. (1997). *Cohort study of learner and novice drivers: Part 4. Novice driver accidents in relation to methods of learning to drive, performance in the driving test and self assessed driving ability and behaviour (*TRL Report 275). Crowthorne, UK: Transport Research Laboratory.

Maycock, G., & Lockwood, C.R. (1993). The accident liability of British car drivers. *Transport Reviews, 13*, 231–245.

Mayhew, D.R. (2007). Driver education and graduated licensing in North America: Past, present, and future. *Journal of Safety Research, 38*, 229–235.

Mayhew, D.R., & Simpson, H.M. (2002). The safety value of driver education and training. *Injury Prevention*, *8*, ii3–ii8.

Mayhew, D.R., Simpson, H.M., & Pak, A. (2003). Changes in collision rates among novice drivers during the first months of driving. *Accident Analysis & Prevention, 35*, 683–691.

Mayhew, D.R., Simpson, H.M., Williams, A.F., & Ferguson, S.A. (1998). Effectiveness and role of driver education and training in a graduated licensing system. *Journal of Public Health Policy, 19*, 51–67.

McCall, J.C., Achler, O., & Trivedi, M.M. (2004). Design of an instrumented vehicle test bed for developing a human centered driver support system. *Proceedings of the IEEE Intelligent Vehicles Symposium*, Parma, Italy, 483–488.

McCauley, M.E. (2006). *Do army helicopter training simulators need motion bases?* (Technical Report 1176). Arlington, VA: U.S. Army Research Institute for the Behavioral & Social Sciences.

McDonald, R.P., & Mulaik, S.A. (1979). Determinacy of common factors: A nontechnical review. *Psychological Bulletin, 86*, 297–306.

McKnight, A.J., & Adams, B.B. (1970). *Driver education task analysis. Volume II: task analysis methods* (Report. DOT-HS-800-368). Washington, DC: National Highway Traffic Safety Administration.

McKnight, A.J., & McKnight, A.S. (2003). Young novice drivers: Careless or clueless? *Accident Analysis & Prevention, 35*, 921–925.

McRuer, D.T., Allen, R.W., Weir, D.H., & Klein, R.H. (1977). New results in driver steering control models. *Human Factors, 19*, 381–397.

McRuer, D.T., & Jex, H.R. (1967). A review of quasi-linear pilot models. *IEEE Transactions on Human Factors in Electronics, HFE-8*, 231–249.

Menaker, E.S., & Coleman, S.L. (2007). Learning Styles Again: Where is Empirical Evidence? *Proceedings of the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC)*, Orlando, FL.

Merrill, M.D. (2002). First principles of instruction. *Educational Technology Research and Development, 50*, 43–59.

Merritt, J.O., & CuQlock-Knopp, V.G. (1991). Perceptual training with cues for hazard detection in off-road driving. *Proceedings of the SPIE*, 133–138.

Mesken, J. (2006). *Determinants and consequences of drivers' emotions.* Doctoral dissertation, University of Groningen, the Netherlands.

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

Michon, J.A. (1985). A critical view of driver behavior models: What do we know, what should we do? In L. Evans & R.C. Schwing (Eds.), *Human behavior and traffic safety* (pp. 485–520). New York: Plenum.

Michon, J.A. (Ed.). (1993). *Generic Intelligent Driver Support: A comprehensive report on GIDS*. London: Taylor & Francis.

Mollenhauer, M.A. (2004). *Simulator adaptation syndrome literature review*. Royal Oak, MI: Realtime Technologies.

Mollenhauer, M.A., Romano R.A., & Brumm, B. (2004). The evaluation of a motion base driving simulator in a CAVE at TACOM. *Proceedings of the 24th Army Science Conference*, Orlando, FL.

Mourant, R.R., & Parsi, L. (2002). Training in a virtual stereoscopic environment. *Proceedings of the Human Factors Society 46th Annual Meeting*, Baltimore, MD, 2206–2209.

Mowafy, L., & Thurman, R.A. (1993). Training pilots to visualize large-scale spatial relationships in a stereoscopic display. *Proceedings of the SPIE, 1915*, 72–81.

Mulder, M. (1999). *Cybernetics of tunnel-in-the-sky displays*. Doctoral dissertation, Delft University of Technology, the Netherlands.

Mulder, M., Mulder, M., Van Paassen, M.M., & Abbink, D.A. (2005). Effects of lead vehicle speed and separation distance on driver car-following behavior. *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, 399–404.

Mundfrom, D.J., Shaw, D.G., & Ke, T.L. (2005). Minimum sample size recommendations for conducting factor analyses. *International Journal of Testing, 5*, 159–168.

Näätänen, R., & Summala, H. (1974). A model for the role of motivational factors in drivers' decision-making. *Accident Analysis & Prevention, 6*, 243–261.

Nap, R.C., De Winter, J.C.F., & Wieringa, P.A. (2007). A method for modelling operator behaviour applied to the NewTranspall microworld. *Proceedings of the 27th European Annual Conference on Human Decision Making and Manual Control*, Lyngby, Denmark.

Nash, E.B., Edwards, G.W., Thompson, J.A., & Barfield, W. (2000). A review of presence and performance in virtual environments. *International Journal of Human-Computer Interaction, 12*, 1–41.

Neerincx, M.A., Lindenberg, J., & Grootjen, M. (2005). Accessibility on the job: Cognitive capacity driven personalization. *Proceedings HCII*, St. Louis, MO: MIRA Digital Publishing.

Neisser, U., Boodoo, G., Bouchard, T.J., Jr., Boykin, A.W., Brody, N., Ceci, S.J., et al. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51*, 77–101.

Newell, A. (1992). Unified theories of cognition, In J.A. Michon & A. Akyürek (Eds.), *Soar: A cognitive architecture in perspective*. Cambridge, MA: Kluwer Academic.

Nordmark, S., Jansson, H., Palmkvist, G., & Sehammar, H. (2004). The new VTI simulator. Multipurpose moving base with high performance linear motion. *Proceedings of the Driving Simulation Conference Europe*, Paris, France, 45–55.

Norman, D.A., & Shallice, T. (1986). Attention to action: Willed and automatic control of behaviour. In R.J. Davidson, G.E. Schwartz, & D. Shapiro (Eds.), *Consciousness and self-regulation: Advances in research and theory* (pp. 1–18). New York: Plenum Press.

Nyberg, A. (2007). *The potential of driver education to reduce traffic crashes involving young drivers.* Doctoral dissertation. Linköping University, Sweden.

Nyberg, A., Gregersen, N.P., & Wiklund, M. (2007). Practicing in relation to the outcome of the driving test. *Accident Analysis & Prevention, 39*, 159–168.

O'Connor, B.P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instruments, & Computers, 32*, 396–402.

O'Neill, B. (1977). A decision-theory model of danger compensation. *Accident Analysis & Prevention, 9*, 157–165.

O'Neill, B., & Williams, A. (2004). Risk homeostasis hypothesis: A rebuttal. *Injury Prevention, 4*, 92–93.

Oliver, M. (2005). The problem with affordance. *E-Learning, 2*, 402–413.

Organisation for Economic Co-operation and Development (2006). *Young drivers: The road to safety.* Paris, France: OECD.

Östlund, J., Peters, B., Thorslund, B., Engström, J., Markkula, G., Keinath, A., et al. (2004). *Driving performance assessment methods and metrics* (AIDE D2.2.5). European Commission.

Özkan, T., & Lajunen, T. (2005). A new addition to DBQ: Positive Driver Behaviours Scale. *Transportation Research Part F: Traffic Psychology and Behaviour, 8*, 355–368.

Page, Y., Ouimet, M.C., & Cuny, S. (2004). An evaluation of the effectiveness of the supervised driver-training system in France. *Annual Proceedings of the Association for the Advancement of Automotive Medicine, 48*, 131–145.

Panou, M., & Bekiaris, E. (2007). TRAIN-ALL: Integrated systems for driver training and assessment using interactive education tools and new training curricula for all modes of transport. Retrieved 10 July 2008 from http://www.trainall-eu.org

Panou, M., Bekiaris, E., & Papakostopoulos, V. (2007). Modelling driver behaviour in European Union and International Projects. In P.C. Cacciabue (Ed.), *Modelling driver behaviour in automotive environments: Critical issues in driver interactions with intelligent transport systems* (pp. 3–25). London: Springer-Verlag.

Parker, D. (2007). Driver error and crashes. In P.C. Cacciabue (Ed.), *Modelling driver behaviour in automotive environments: Critical issues in driver interactions with intelligent transport systems* (pp. 266–274). London: Springer-Verlag.

Parker, D., Lajunen, T., & Stradling, S. (1998). Attitudinal predictors of interpersonally aggressive violations on the road. *Transportation Research Part F: Traffic Psychology and Behaviour, 1*, 11–24.

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

Parkes, A.M. (2005). Improved realism and improved utility of driving simulators: Are they mutually exclusive? *HUMANIST Workshop on the application of new technologies to driver training*, Brno, Czech Republic. Retrieved 10 July 2008 from http://www.esafetysupport.org/download/research_and_development/HUMANISTA_05Improved.pdf

Peden, M., Scurfield, R., Sleet, D., Mohan, D., Hyder, A.A., Jarawan, E., et al. (2004). *World report on road traffic injury prevention*. Geneva, Switzerland: World Health Organization.

Pennell, R. (1968). The influence of communality and *N* on the sampling distributions of factor loadings. *Psychometrika, 33*, 423–439.

Pfautz, J.D. (2001). Sampling artifacts in perspective and stereo displays. *Proceedings of the SPIE*, *4297*, 54–62. Retrieved 16 July 2008 from http://www.mit.edu/~jpfautz/spie-sda-4297a-08.pdf

Piao, J., & McDonald, M. (2003). Low speed car following behaviour from floating vehicle data. *Proceedings of the IEEE Intelligent Vehicles Symposium*, Columbus, OH, 462–467.

Pinto, M., Cavallo, V., Ohlmann, T., Espié, S., & Roge, J. (2004). The perception of longitudinal accelerations: What factors influence braking manoeuvers in driving simulators. *Proceedings of the Driving Simulation Conference Europe*, Paris, France, 139–151.

Pirenne, D., Arno, P., Baten, G., & Breker, S. (2002). *TRAINER: Assessment criteria and methodology* (Deliverable 5.1). Brussels, Belgium: European Commission.

Pitt, M.A., Young, I.J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review, 109*, 472–491.

Plöchl, M., & Edelmann, J. (2007). Driver models in automobile dynamics application. *Vehicle System Dynamics, 45*, 699–741.

Polet, P., Vanderhaegen, F., & Wieringa, P.A. (2002). Theory of safety-related violations of system barriers. *Cognition, Technology & Work, 4*, 171–179.

Preacher, K.J. (2003). *The role of model complexity in the evaluation of structural equation models*. Doctoral dissertation. Ohio State University.

Preacher, K.J., & MacCallum, R.C. (2002). Exploratory factor analysis in behavior genetics research: Factor recovery with small sample sizes. *Behavior Genetics, 32*, 153–161.

Quenault, S.W., Golby, C.W., & Pryer, P.M. (1968). *Age group and accident rate – driving behaviour and attitudes* (Report LR167). Crowthorne, UK: Transportation and Road Research Laboratory.

Ranney, T.A. (1994). Models of driving behavior: A review of their evolution. *Accident Analysis & Prevention, 26*, 733–750.

Ranney, T.A. (1999). Psychological factors that influence car-following and car-following model development. *Transportation Research Part F: Traffic Psychology and Behaviour, 2*, 213–219.

Rasmussen, J. (1983). Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. *IEEE Transactions on Systems, Man, and Cybernetics, 13*, 257–266.

Reason, J.T. (1999). *Managing the risks of organizational accidents*. Aldershot: Ashgate Publishing Company.

Reason, J.T., Manstead, A., Stradling, S., Baxter, J., & Campbell, K. (1990). Errors and violations on the roads: A real distinction? *Ergonomics, 33*, 1315–1332.

Reed, M.P., & Green, P.A. (1999). Comparison of driving performance on-road and in a low-cost simulator using a concurrent telephone dialling task. *Ergonomics, 42*, 1015–1037.

Reid, L.D. (1983). A survey of recent driver steering behavior models suited to accident studies. *Accident Analysis & Prevention, 15*, 23–40.

Reymond, G., Kemeny, A., Droulez, J., & Berthoz, A. (2001). Role of lateral acceleration in curve driving: Driver model and experiments on a real vehicle and a driving simulator. *Human Factors, 43*, 483–495.

Riecke, B.E., Schulte-Pelkum, J., Caniard, F., & Bülthoff, H.H. (2005). Towards lean and elegant self-motion simulation in virtual reality. *Proceedings of IEEE Virtual Reality 2005*, 131–138.

Rimmö, P-A., & Åberg, L. (1999). On the distinction between violations and errors: Sensation seeking associations. *Transportation Research Part F: Traffic Psychology and Behaviour*, *2*, 151–166.

Robertson, D., & Symons, J. (2007). Maximum likelihood factor analysis with rank-deficient sample covariance matrices. *Journal of Multivariate Analysis, 98*, 813–828.

Roenker, D.L., Cissell, G.M., Ball, K.K., Wadley, V.G., & Edwards, J.D. (2003). Speed-of-processing and driving simulator training result in improved driving performance. *Human Factors, 45*, 218–233.

Rothengatter, T. (1988). Risk and the absence of pleasure: A motivational approach to modelling road user behaviour. *Ergonomics, 31*, 599–607.

Rothengatter, T. (1991). Automatic policing and information systems for increasing traffic law compliance. *Journal of Applied Behavior Analysis, 24*, 85–87.

Rothengatter, T. (1997). Errors and violations as factors in accident causation. In T. Rothengatter & E. Carbonell Vaya (Eds.), *Traffic and Transport Psychology, Theory and Application* (pp. 59–64). Amsterdam: Pergamon.

Rothengatter, T. (2002). Drivers' illusions – no more risk. *Transportation Research Part F: Traffic Psychology and Behaviour, 5*, 249–258.

Roza, Z.C. (2005). *Simulation fidelity theory and practice: A unified approach to defining, specifying and measuring the realism of simulations*. Doctoral dissertation, Delft University of Technology, the Netherlands.

Sachsenweger, M., & Sachsenweger U. (1991). Stereoscopic acuity in ocular pursuit of moving objects. *Documenta Ophthalmologica, 78*, 7–128.

Salas, E., Bowers, C.A., & Rhodenizer, L. (1998). It is not how much you have but how you use it: Toward a rational use of simulation to support aviation training. *The International Journal of Aviation Psychology, 8*, 197–208.

Salthouse, T.A. (1979). Adult age and the speed-accuracy trade-off. *Ergonomics, 22*, 811–821.

Sapnas, K.G., & Zeller, R.A. (2002). Minimizing sample size when using exploratory factor analysis for measurement. *Journal of Nursing Measurement, 10*, 135–154.

Sato, T., & Akamatsu, M. (2008). Modeling and prediction of driver preparations for making a right turn based on vehicle velocity and traffic conditions while approaching an intersection. *Transportation Research Part F: Traffic Psychology and Behaviour, 11*, 242–258.

Schönemann, P.H. (1990). Facts, fictions, and common sense about factors and components. *Multivariate Behavioral Research, 25*, 47–51.

Schwebel, D.C., Severson, J., Ball, K.K., & Rizzo, M. (2006). Individual difference factors in risky driving: The roles of anger/hostility, conscientiousness, and sensation-seeking. *Accident Analysis & Prevention, 38*, 801–810.

Senserrick, T.M. (2001). New look driver-training: Deflating confidence and promoting safety. *Paper presented to the Road Safety Research, Policing and Education Conference,* Melbourne, Australia. Retrieved 10 July 2008 from http://www.monash.edu.au/cemo/roadsafety/abstracts_and_papers/025/SENSERR1.pdf

Senserrick, T.M., Brown, T., Marshall, D., Quistberg, D.A., Dow, B., & Winston, F.K. (2007). Risky driving by recently licensed teens: Self-reports and simulated performance. *Proceedings of the Fourth International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*, Stevenson, WA, 541–548.

Senserrick, T.M., & Haworth, N. (2005). *Review of literature regarding national and international young driver training, licensing and regulatory systems* (Report No. 239). Clayton, Australia: Monash University, Accident Research Centre.

Showalter, T.W., & Parris, B.L. (1980). *The effects of motion and g-seat cues on pilot simulator performance of three piloting task* (TP-1601). NASA.

Siegler, I., Reymond, G., Kemeny, A., & Berthoz, A. (2001). Sensorimotor integration in a driving simulator: Contributions of motion cueing in elementary driving tasks. *Proceedings of the Driving Simulation Conference*, Sophia-Antipolis, France, 21–32.

Simons-Morton, B., & Ouimet, M.C. (2006). Parent involvement in novice teen driving: A review of the literature. *Injury Prevention, 12*, i30–i37.

Spiers, H.J., & Maguire, E.A. (2007). Neural substrates of driving behaviour. *NeuroImage, 36*, 245–255.

Spiker, V.A., Karp, M.R., Mautone, T., & Fischer, S. (2007). Do better multi-taskers make better pilots? *Proceedings of the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC)*, Orlando, FL.

Srivastava, S., John, O.P., Gosling, S.D., & Potter, J. (2003). Development of personality in early and middle adulthood: Set like plaster or persistent change? *Journal of Personality and Social Psychology, 84*, 1041–1053.

Stanton, N.A., Walker, G.H., Young, M.S., Kazi, T., & Salmon, P.M. (2007). Changing drivers' minds: The evaluation of an advanced driver coaching system. *Ergonomics, 50*, 1209–1234.

Stanton, N.A., Young, M., & McCaulder, B. (1997). Drive-by-wire: The case of driver workload and reclaiming control with adaptive cruise control. *Safety Science, 27*, 149–159.

Steiger, J.H. (1994). Factor analysis in the 1980‘s and the 1990‘s: Some old debates and some new developments. In I. Borg & P. Ph. Mohler (Eds.), *Trends and Perspectives in Empirical Social Research* (pp. 201–224). Berlin: Walter de Gruyter.

Stelmach, G.E., & Nahom, A. (1992). Cognitive-motor abilities of the elderly driver. *Human Factors, 34*, 53–65.

Stewart, A.E., & St. Peter, C.C. (2004). Driving and riding avoidance following motor vehicle crashes in a non-clinical sample: Psychometric properties of a new measure. *Behaviour Research and Therapy, 42*, 859–879.

Strayer, D.L., Cooper, J.M., & Drews, F.A. (2008). *Part-task and variable priority simulator-based training for snowplow operators*. Manuscript submitted for publication.

Strayer, D.L., & Drews, F.A. (2003). Simulator training improves driver efficiency: Transfer from the simulator to the real world. *Proceedings of the Second International Driving*

Symposium on Human Factors in Driver Assessment, Training and Vehicle Design, Park City, UT, 190–193.

Strayer, D.L., Drews, F.A., & Burns, S. (2005). The development and evaluation of a high-fidelity simulator training program for snowplow operators. *Proceedings of the Third International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*, Rockport, ME, 464–470.

Summala, H. (1996). Accident risk and driver behaviour. *Safety Science, 22*, 103–117.

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

Suzuki, K., & Jansson, H. (2003). An analysis of driver's steering behaviour during auditory or haptic warnings for the designing of lane departure warning system. *JSAE Review, 24*, 65–70.

SWOV Institute for Road Safety Research (2006). *SWOV Fact sheet. Simulators in driver training*. Retrieved 10 July 2008 from http://www.swov.nl/rapport/Factsheets/UK/FS_Simulators_in_driver_training.pdf

SWOV Institute for Road Safety Research (2008). *Risk – victims per 100000 inhabitants. Source: Statistics Netherlands / Ministry of Transport*. Retrieved 10 July 2008 from http://www.swov.nl/cognos/cgi-bin/ppdscgi.exe.

Tabachnick, B.G., & Fidell, L.S. (2001). *Using multivariate statistics*. Boston, MA: Allyn & Bacon.

Theeuwes, J. (2001). The effects of road design on driving. In P.-E. Barjonet (Ed.), *Traffic psychology today* (pp. 241–264). Boston: Kluwer Academic Publishers.

Tidwell, R.P. (1990). Stereopsis takes off in flight simulation. *Proceedings of the IEEE SoutheastCon*, New Orleans, LA, 578–582.

Tomaske, W., Breidenbach, C., & Fortmüller, T. (2001). A scientific and physiological research study with truck driving simulators in the army. *Proceedings of the International Training & Education Conference (ITEC)*, Lille, France.

Toroyan, T., & Peden, M. (2007). *Youth and road safety*, Geneva, Switzerland: World Health Organization.

Trimpop, R.M. (1996). Risk homeostasis theory: Problems of the past and promises for the future. *Safety Science, 22*, 119–130.

Tucker, L.R., Koopman, R.F., & Linn, R.L. (1969). Evaluation of factor analytic research procedures by means of simulated correlation matrices. *Psychometrika, 34*, 421–459.

Turpin, D.R., Welles, R.T., & Price, C. (2007). Simulator-based learning: Achieving performance improvement independent of instructors. *Proceedings of the Fourth International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*, Stevenson, WA, 481–487.

Twisk, D.A.M., & Stacey, C. (2007). Trends in young driver risk and countermeasures in European countries. *Journal of Safety Research, 38*, 245–257.

Uhr, M.B.F., Felix, D., Williams, B.J., & Krueger, H. (2003). Transfer of training in an advanced driving simulator: Comparison between real world environment and simulation in a manoeuvering driving task. *Proceedings of the Driving Simulation Conference North America*, Dearborn, MI.

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

Vaa, T. (2007). Modelling driver behaviour on basis of emotions and feelings: Intelligent transport systems and behavioural adaptations. In P.C. Cacciabue (Ed.), *Modelling* *driver behaviour in automotive environments: Critical issues in driver interactions with intelligent transport systems* (pp. 208–232). London: Springer-Verlag.

Vaden, E.A., & Hall, S. (2005). The effect of simulator platform motion on pilot training transfer: A meta-analysis. *The International Journal of Aviation Psychology, 15*, 375–393.

Van der Molen, H.H., & Bötticher, A.M.T. (1988). A hierarchical risk model for traffic participants. *Ergonomics, 31*, 537–555.

Van der Snee, E. (2005). *Evaluatie gebruik De Nederlandse Rijsimulator bij rijscholen*. Unpublished report, Delft University of Technology, the Netherlands.

Van Emmerik, M.L. (2004). *Beyond the simulator: Instruction for high-performance tasks.* Doctoral dissertation, University of Twente, the Netherlands.

Van Winsum, W. (1999). The human element in car following models. *Transportation Research Part F: Traffic Psychology and Behaviour, 2*, 207–211.

Van Winsum, W., & Godthelp, H. (1996). Speed choice and steering behaviour in curve driving. *Human Factors*, *38*, 434–441.

Vargas-Martín, F., & García-Pérez, M.A. (2005). Visual fields at the wheel. *Optometry and Vision Science, 82*, 675–681.

Velicer, W.F., & Fava, J.L. (1998). Effects of variable and subject sampling on factor pattern recovery. *Psychological Methods, 3*, 231–251.

Velicer, W.F., & Jackson, D.N. (1990). Component analysis versus common factor analysis: Some further observations. *Multivariate Behavioral Research, 25*, 97–114.

Vernick, J.S., Li, G., Ogaitis, S., MacKenzie, E.J., Baker, S.P., & Gielen, A.C. (1999). Effects of high school driver education on motor vehicle crashes, violations, and licensure. *American Journal of Preventive Medicine, 16*, 40–46.

Verstegen, D.M.L. (2003). *Iteration in instructional design: An empirical study on the specification of training simulators.* Doctoral dissertation, Utrecht University, the Netherlands.

Vicente, K.J. (1999). *Cognitive work analysis: Toward safe, productive, and healthy computer-based work.* Mahwah, NJ: Lawrence Erlbaum Associates.

Vissers, J.A.M.M. (2001). *Het Nieuw Rijden. Handleiding voor de rijopleiding B* (Report No. 4HNR-01.10). Novem.

Vlakveld, W.P. (2000). *Leerdoelen voor het rijbewijs B*. Rotterdam, the Netherlands: Rijkswaterstaat. Adviesdienst Verkeer en Vervoer, Ministerie van Verkeer en Vervoer.

Vlakveld, W.P. (2005a). *Jonge beginnende automobilisten, hun ongevalsrisico en maatregelen om dit terug te dringen: Een literatuurstudie [Young, novice motorists, their crash rates, and measures to reduce them: A literature study]*. (Report No. R-2005-3). Leidschendam, the Netherlands: SWOV.

Vlakveld, W.P. (2005b). The use of simulators in basic driver training. *HUMANIST Workshop on the application of new technologies to driver training*, Brno, Czech Republic. Retrieved 10 July 2008 from http://www.esafetysupport.org/download/research_and_development/HUMANISTA_13Use.pdf

Vlakveld, W.P. (2006a). *Veiligheidswaarde van de ANWB-rijopleiding* (Report No. D-2006-5). Leidschendam, the Netherlands: SWOV.

Vlakveld, W.P. (2006b). Will simulator training in basic driver education help to enhance road safety? *HUMANIST Workshop on the application of new technologies to driver training*, Madrid, Spain. Retrieved 10 July 2008 from http://www.noehumanist.org/documents/Madrid_2006-24/12-madrid_Paper4-3_Vlakveld.pdf

Vlakveld, W.P. (2008). Hazard perception test (Virtual reality). Retrieved 25 August 2008 from http://www.cieca.be/download/Vlakveld.pdf

Von der Heyde, M., & Riecke, B.E. (2001). *How to cheat in motion simulation – comparing the engineering and fun ride approach to motion cueing* (Report No. 089). Tübingen, Germany: Max Planck Institute for Biological Cybernetics.

Wallén Warner, H., & Åberg, L. (2006). Drivers' decision to speed: A study inspired by the theory of planned behavior. *Transportation Research Part F: Traffic Psychology and Behaviour, 9*, 427–433.

Wassink, I., Van Dijk, B., Zwiers, J., Nijholt, A., Kuipers, J., & Brugman, A. (2006). In the Truman Show: Generating dynamic scenarios in a driving simulator. *IEEE Intelligent Systems, 21*, 28–32.

Weevers, I., Kuipers, J., Brugman, A.O., Zwiers, J., Van Dijk, E.M.A.G., & Nijholt, A. (2003a). The virtual driving instructor: Creating awareness in a multiagent system. In Y. Xiang & B. Chaib-draa (Eds.), *Advances in Artificial Intelligence: 16th Conference of the Canadian Society for Computational Studies of Intelligence, Halifax, Canada* (pp. 596–602). Berlin Heidelberg: Springer-Verlag.

Weevers, I., Nijholt, A., Kuipers, J., Van Dijk, E.M.A.G., Brugman, A.O., & Zwiers, J. (2003b). A student-adaptive system for driving simulation. *Proceedings of the 8th Australian and New Zealand Intelligent Information Systems Conference*, Sydney, Australia, 253–257.

Welles, R.T., & Holdsworth, M. (2000). Tactical driver training using simulation: Recent experiences in law enforcement driving simulation. *Proceedings of the Interservice/Industry Training, Simulation & Education Conference (I/ITSEC)*, Orlando, FL.

Wells, P., Tong, S., Sexton, B., Grayson, G., & Jones, E. (2008). *Cohort II: A study of learner and new drivers: Volume 1 – Main Report* (No. 81). London, UK: Department for Transport.

Westlake, W. (2001). Is a one eyed racing driver safe to compete? Formula one (eye) or two? *British Journal of Ophthalmology, 85*, 619–624.

Wierda, M. (1996). Leren rijden zonder auto. In F.J.J.M. Steyvers & P.G.M. Miltenburg (Eds.), *Gedragsbeïnvloeding in verkeers- en vervoerbeleid* (pp. 25–29), Rijksuniversiteit Groningen.

Wheeler, W.A., & Trigs, T.J. (1996). A task analytical view of simulator based training for drivers. *Proceedings of the road safety research and enforcement conference 'effective partnerships'*, Coogee Beach, New South Wales, Australia, 217–221.

Whitelock, D., Romano, D., Jelfs, A., & Brna, P. (2000). Perfect presence: What does this mean for the design of virtual learning environments? *Education and Information Technologies, 5*, 277–289.

Wiberg, M. (2006). Gender differences in the Swedish driving-license test. *Journal of Safety Research, 37*, 285–291.

Wiesenthal, D.L., Hennessy, D., & Gibson, P.M. (2000). The Driving Vengeance Questionnaire (DVQ): The development of a scale to measure deviant drivers' attitudes. *Violence and Victims, 15*, 115–136.

Wilde, G.J.S. (1988). Risk homeostasis theory and traffic accidents: Propositions, deductions and discussion of dissension in recent reactions. *Ergonomics, 31*, 441–468.

Wilde, G.J.S., & Robertson, L.S., & Pless, I.B. (2002). For and against: Does risk homoeostasis theory have implications for road safety. *British Medical Journal, 324*, 1149–1152.

Williams, A.F. (2003). Teenage drivers: Patterns of risk. *Journal of Safety Research, 34*, 5–15.

Williams, A.F. (2006). Young driver risk factors: Successful and unsuccessful approaches for dealing with them and an agenda for the future. *Injury Prevention, 12*, i4–i8.

Williams, A.F., & O'Neill, B. (1974). On-the-road driving records of licensed race drivers. *Accident Analysis & Prevention, 6*, 263–270.

World Health Organization (2002). Gender and road traffic injuries. Geneva, Switzerland. Retrieved 10 July 2008 from http://www.who.int/gender/other_health/en/gendertraffic.pdf

Xie, C., & Parker, D. (2002). A social psychological approach to driving violations in two Chinese cities. *Transportation Research Part F: Traffic Psychology and Behaviour*, *5*, 293–308.

Ximénez, C. (2006). A Monte Carlo study of recovery of weak factor loadings in confirmatory factor analysis. *Structural Equation Modeling, 13*, 587–614.

Yan, X., Abdel-Aty, M., Radwan, E., Wang, X., & Chilakapati, P. (2008). Validating a driving simulator using surrogate safety measures. *Accident Analysis & Prevention, 40*, 274–288.

Yuasa, S., Kurachi, M., Suzuki, M., Kadono, Y., Matsui, M., Saitoh, O., & Seto, H. (1995). Clinical symptoms and regional cerebral blood flow in schizophrenia. *European Archives of Psychiatry and Clinical Neuroscience, 246*, 7–12.

Zaroff, C.M., Knutelska, M., & Frumkes, T.E. (2003). Variation in stereoacuity: Normative description, fixation disparity, and the roles of aging and gender. *Investigative Ophthalmology and Visual Science, 44*, 891–900.

Zaychik, K.B., & Cardullo, F.M. (2003). Simulator sickness: The problem remains. *AIAA Modeling and Simulation Technologies Conference and Exhibit*, Austin, TX, 11–14.

Zeller, R.A. (2006). *Statistical tools in applied research*. Retrieved 10 July 2008 from http://www.personal.kent.edu/~rzeller/Ch.%2010.pdf

Zhai, S., Accot, J., & Woltjer, R. (2004). Human action laws in electronic virtual worlds: An empirical study of path steering performance in VR. *Presence: Teleoperators and Virtual Environments, 13*, 113–127.

Zhu, M., & Ghodsi, A. (2006). Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Computational Statistics and Data Analysis, 51*, 918–930.

Acknowledgements

Despite the single name on the cover, a PhD thesis is never completed in isolation. I would like to thank a number of people for their contribution to this research.

Jorrit Kuipers, director of Green Dino Virtual Realities (GDVR), is one of the initiators of this project. From early on, Jorrit has recognized the importance of research and data storage facilities in driving simulation. I would like to thank Jorrit for his knowledge, enthusiasm, and innovative mind-set. I would like to use this opportunity to also thank GDVR team members who I had contact with, including Andrea Poelstra, Arnd Brugman, Olaf Achthoven, Gijs Harmens, and Bartjan Engelfriet.

Prof. Peter Wieringa was a leading adviser in this project. I would like to thank Peter for his coaching, his expertise in human-machine systems, and for involving me in various educational activities. Prof. Bob Mulder showed great involvement and support as well. His knowledge on control and simulation and his "mathematical" focus have been very helpful.

I would like to express my gratitude to other direct mentors. Max Mulder has been extensively involved as an adviser and has provided imperative know-how about human-machine interaction. René van Paassen has made important contributions as well, particularly regarding control theory matters. The expertise of Jenny Dankelman on human-machine interaction in the medical domain has been indispensable.

I had a synergistic cooperation with my colleagues in the project. Harm Boschloo has provided important solutions regarding software programming and the implementation of the vibrating seat. Stefan de Groot, with his excellent research attitude, has provided important insight regarding augmented feedback and instructions in simulation-based training. Stefan and I had a good cooperation and many useful discussions. Without his efforts, this thesis would not be as complete as it is now.

This research was subsidized by the Dutch Ministry of Economic Affairs, under its innovation research programme on man-machine interaction (IOP MMI). I would like to express gratitude to the following programme committee members for their commitment and for facilitating this opportunity for research: René Collier, Martijn Nuijten, and Hans Kruithof. In addition, I express my appreciation towards the coaching committee members, who have provided important suggestions during half-yearly meetings: Jelke van der Pal, Lucas Noldus, Mark Neerincx, Ad van Lier, Karin Schoneveld, Peter Oosterpoort, and Rob van Egmond. Willem Vlakveld deserves to be mentioned separately for his critique and suggestions on a number of chapters.

The Dutch Driving Test Organisation (CBR) is very much acknowledged for the interest and involvement in our research and for providing us with useful data. In particular, I would like to thank Theo van Rijt and Piet Fuchs.

In the project, we have guided a number of undergraduate students who have worked in teams on various research assignments. Their results have provided a major momentum to this thesis both by means of discussions and by generating experimental data. A number of these contributions have culminated in scientific publications (in particular chapters 7, 8, and 9). I would like to thank the following persons for their excellent work, including the construction of several of the motion cueing systems, and for long laboratory hours: Krijn van Aken, Chris Arentsen, Birger Jansen, Martijn Lindhout, René van Duijn, Marijke van den Berg, Joost van Gijn, Tim ter Horst, Linh Ta Cam, Wouter Lodewijk, Najim Mahdaoui, Eva Promes, Kai Balder, Casper Louw, Joep Mutsaerts, Douwe Starink, David Epema, Sven Molenaar, Rudi van der Sar, Radjen Vervuurt, Dominique Biever, Keith Gonesh, Pieter van der Meer, Jasper Winters, Henri Boessenkool, Jim de Clercq, Ron Kaandorp, Ard Veldhuijzen, Eric Haardt, Kai van Rhede van der Kloot, Alexander Zonneveld, Berend de Graaf, Benno Poppelaars, Duco Kaasjager, Jasper Truffino, Pieter van Gaelen, Kirsten Henken, Niels Kuijken, Boudewijn Visser, Rik Doorten, Ruben Griffioen, Hugo Reijkens, Florian Wasser, Hoai van Huynh, Siamak Mohammadian, Andrew Sidharta, Paul van der Ploeg, Christiaan Rademakers, Daphne Swemmers, Phu Do, Arjan van Leeuwen, Stefan van Loenhout, Vincent Gusdorf, Roel van Gorkum, Maarten Smit, Riccardo van der Ende, Kevin Bohnen, Arthur Fassotte, Corné Groen, Rob Vellekoop, Pascal ‘t Hart, Jurriaan Knobel, Arjan van der Raad, and Bart de Vette.

Two internship students from The Hague University of Applied Sciences have contributed in this project. Elske van der Snee conducted interviews with driving school owners and students, thereby generating valuable data. Richard Fickert made an important contribution to the graphical design of the strength-weakness report, currently used in Dutch Driving Simulators.

I had the privilege of advising two MSc students during their research; their results have provided further stimulation to the ideas in this thesis. Robert Nap conducted fundamental research on intentional violations of operators in virtual environments. Daan Venekamp performed research into data mining large amounts of hospital event reports.

During the past years, we worked together with researchers from Université de Valenciennes of the Laboratoire d’Automatique, de Mécanique et d’Informatique industrielles et Humaines (LAMIH). I would like to thank Frédéric Vanderhaegen for his enthusiasm, his experience on human behaviour in driving simulators, and for involving me in a European network of researchers on human-machine interaction.

In the final stage, I had important help from my father-in-law Frans Nieuwenhof, who spent many hours formatting the thesis. An experienced scientist himself, Frans has also been able to comment on the content of this work. Without his major editing efforts, the thesis would not be in the shape it is now. In addition, Dimitra Dodou has helped editing the figures. Being a competent researcher, she has also provided useful insight and engaged in fundamental discussions in many stages, which has

also lead to a joint publication with practically equal amount of effort spent (Appendix B).

I would like to thank Riender Happee for providing useful feedback on the manuscript. Erwin Boer has provided important suggestions on several occasions. David Abbink, Mark Mulder, Magda Chmarra, and Stefan Klein are acknowledged for useful discussions and joint contributions on simulation and training.

I would like to thank Maura Houtenbos for running two experiments in our lab, and for providing access to the raw data. I would also like to thank her supporting team for pleasant and useful meetings: Andrew Hale, Marjan Hagenzieker, and Tom Heijer. In addition, Aart Spek and Arnaud Herbelin have run an experiment in our simulator; this experiment has provided useful insight and has resulted in a joint research paper.

Special thanks go to Maarten Vriezen and Esther Blom for doing an exploratory study on low-cost motion cueing solutions. Jeroen den Dekker is a good friend and colleague on human-machine interfaces in transportation and creator of the cover of this thesis.

I am grateful to Dr. Samuel D. Gosling for providing the dataset of the Big Five Inventory for the subsampling study included in Appendix B.

I would like to thank Frans van der Meijden and John Seiffers, as well as many others of the TU Delft staff for helping out in actuator control problems and for providing access to sensory equipment and other hardware. I would also like to thank the colleagues who showed interest and supported our work.

Finally I thank a number of persons in my close environment: My parents and my sisters Anne en Karlijn for their interest and continuing support. During the course of this PhD study, our children Jet and Arthur have been born, making me aware that there are more important things in life than completing a thesis. Last but not least, my wife Janneke for support, understanding, and great sandwiches!

Curriculum vitae

7 March 1979
Born in Utrecht, the Netherlands.

1997 – 2004
MSc study AeroSpace Engineering at Delft University of Technology. Graduated at the Control and Simulation Department.

2004 – 2009
PhD study at Delft University of Technology, Faculty of Mechanical, Maritime and Materials Engineering, Department of BioMechanical Engineering.