Linköping Studies in Science and Technology
Dissertations, No. 1250

Flight Simulator Training: Assessing the Potential

Staffan Nählinder

Linköping University
INSTITUTE OF TECHNOLOGY

2009

Department of Management and Engineering
Linköpings universitet, SE-581 83 Linköping, Sweden

"Flight Simulator Training: Assessing the Potential"

Linköping Studies in Science and Technology, Dissertation No. 1250

ISBN: 978-91-7393-658-3
ISSN: 0345-7524

Printed by: LiU-Tryck, Linköping

Distributed by:
Linköping University
Department of Management and Engineering
SE-581 83 Linköping, Sweden
Tel: +46 13 281000

ABSTRACT

Mental workload is an important concept and has been proven to be a precursor to situation awareness and operative performance. This thesis describes methods to measure mental workload through self-ratings and psychophysiological measurements. Similarities and differences in psychophysiological reactions and rated mental workload between simulated and real flights are described. The consequences of such similarities and differences are discussed and its possible effect on training potential.

A number of empirical studies are presented. They describe the experience and the psychophysiological reactions of pilots flying in a simulator and in real flight. In most cases, the reactions are similar – there is a high degree of accordance in rated mental workload and psychophysiological reaction between simulated and real flight. The studies show, that even though the responses are similar, there are also interesting differences. In one study, the pilots have consistently lower heart rate, higher heart rate variability and less eye movements in the simulator than in real flight. In another study, during certain events, the pilots have higher heart rate in the simulator than in real flight. The results are important in order to understand the training potential of simulators from a Human Factors perspective. Further, two measurement equipments for psychophysiological recording are compared and various psychophysiological measures are tested in applied settings.

The thesis also discusses some methodological aspects, such as methods to create reliable and valid variables in dynamic applied research and how to deal with individual differences. An algorithm is suggested to remove differences between individuals. This facilitates the finding of within-participant effects.

Finally, results from a study on embedded training tools are presented. In this study, student pilots and instructors rated the usefulness of several embedded training tools. These tools were built into a simulator to facilitate learning and teaching by illustrating concepts that can be difficult to understand. The results show clearly that such training tools are appreciated by both students and instructors. Well implemented, thoroughly selected training tools can dramatically improve the training potential of future training simulators.

SAMMANFATTNING

Mental arbetsbelastning är ett viktigt begrepp som har visat sig kunna predicera bland annat situationsmedvetande och operativ prestation. Avhandlingen visar olika sätt att mäta mental arbetsbelastning, bland annat genom självskattningar och psykofysiologiska mått. Skillnader och likheter i psykofysiologisk reaktion och skattad mental arbetsbelastning mellan simulerad och verklig flygning beskrivs. Betydelsen av sådana skillnader och dess konsekvenser för möjligheten till träningseffekt diskuteras.

Ett antal studier beskrivs som handlar om upplevelsen och de fysiologiska reaktionerna hos piloter som flyger i simulatorer och i verklig flygning. I de flesta fall förekommer likartade reaktioner i simulatorn som i verkligheten. Det finns en stor grad av överensstämmelse både vad gäller psykofysiologisk reaktion och upplevd mental arbetsbelastning. Men studierna visar också att även om reaktionerna är lika, så skiljer de sig också åt på några viktiga punkter. Piloter som genomför ett uppdrag i en simulator är inte lika stressade som i verklig flygning. De har lägre puls och högre pulsvariabilitet. I vissa enstaka fall har piloterna högre puls i simulatorn än i motsvarande fall i verklig flygning. Resultaten är viktiga för att förstå hur nyttan av simulatorer kan utvärderas ur ett användningsperspektiv. Vidare jämförs två olika utrustningar för psykofysiologisk mätning och olika psykofysiologiska mått testas i tillämpade miljöer.

Olika utrustningar för att mäta psykofysiologisk reaktion jämförs och olika psykofysiologiska mått diskuteras. Avhandlingen problematiserar olika metodologiska aspekter, såsom metoder för att skapa reliabla och valida mått i dynamisk tillämpad forskning, samt metoder för att hantera individuella skillnader. En algoritm föreslås för att eliminera olikheter mellan individer. Den underlättar upptäckandet av inomindividseffekter.

Avslutningsvis presenteras resultaten från en studie avsedd att mäta inställning till ett antal inbyggda pedagogiska träningsverktyg. De verktyg som fanns inbyggda i simulatorn var framtagna för att förbättra träningseffekten genom att konkretisera koncept och relationer som kan vara svåra att förstå. Pilotelever och instruktörer fick flyga i en simulator och gavs sedan möjligheten att pröva olika träningsverktyg. Resultaten visar tydligt ett positivt intresse för träningsverktygen både från elever och från instruktörer. Väl implementerade noggrant utvalda träningsverktyg, kan kraftigt förbättra träningseffektiviteten i framtida träningssimulatorer.

ACKNOWLEDGEMENTS

Writing a thesis is never a one mans job. I am extremely grateful for all the support and encouragement I've received from the people around me.

First, I would like to thank my supervisor dr Kjell Ohlsson. Thank you for believing in me and for your encouragement throughout the years. I would also like to thank my "extra" supervisor dr Nils Dahlbäck for his great advice and engagement during the iterative process of this thesis. It would not have been the same without you! My co-supervisor, dr Erland Svensson has always been encouraging and been a great support. He is truly a scientific inspiration to me. My appreciation also includes his wife, Maud Angelborg-Thanderz. You two are both guilty of making me want to work at FOI! I cannot thank you enough.

Elisabeth Peterson, you've made the administrative processes of this thesis work very smoothly indeed, thank you for all your help!

To all the participating pilots at the F17 wing - thank you for participating! Especially thanks to capt Björn Danielsson for making that study possible.

Further, I am in dept to all the kind people and participants at Lund University School of Aviation. Especially thanks to dr Nicklas Dahlström for all your help and push forward and for really making the studies happening! Your insightfulness and kind personality has been imperative. Above all, you are a great friend!

I am also grateful for the overseas collaboration with the USAFRL in Dayton, Ohio. Thank you, dr Glenn Wilson for the work we've done together on psychophysiological measurements. And dr James Christensen, I'm looking forward to continuing this work!

Another great friend and colleague, capt Björn Persson – thank you for all the long and enriching discussions we've had about pedagogics, learning strategies, simulation and training. This is merely the beginning! Let's have sushi.

Thanks Joakim "dr puke" Dahlman, friend and former colleague, for the pep-talks over coffee, pizza and the occasional beer. Keep it real!

My wonderful colleagues and friends at FOI Man-System-Interaction – thank you all for your support, kind words and encouragements along the way! Birgitta, Lars, Martin, Patrik, Peter … you guys are next! Especially I would like to extend my thanks to my great friend Peter Berggren who has always helped me keep a healthy perspective on things.

To my wonderful parents and their belief in me and for all their support, thank you!

Last and most importantly, I thank my wife Johanna for all the love and support you have given me and for our wonderful life together. This thesis would never have been done without you. To Adam and Birger – thank you for putting up with a sometimes distracted daddy. I love you all so much!

LIST OF PUBLICATIONS

This thesis is based on the following six studies. The studies are referred to by their roman numerals. In the year 2002, Magnusson changed his last name to Nählinder.

I Magnusson, S. (2002). Similarities and differences in psychophysiological reactions between simulated and real air-to-ground missions. *International Journal of Aviation Psychology*. Vol. 12(1), pp. 49-61.

II Magnusson, S. & Berggren, P. (2002). Dynamic assessment of pilot mental status. In *Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting* (Baltimore, MD). Human Factors and Ergonomics Society: Santa Monica, CA.

III Dahlström, N. & Nählinder, S. (2006). A comparison of two recorders for obtaining in-flight heart rate data. *Applied Psychophysiology and Biofeedback*, Vol. 31(3), pp. 273-279.

IV Dahlström, N. & Nählinder, S. (2009). Mental workload in simulator and aircraft during basic civil aviation training. *International Journal of Aviation Psychology*. (IN PRESS).

V Dahlström, N., Nählinder, S., Wilson, G. F. & Svensson, E. (2009). Recording of psychophysiological data during aerobatic training. *International Journal of Aviation Psychology*. (ACCEPTED FOR PUBLICATION).

VI Nählinder, S., Berggren, P. & Persson, B. (2005). Increasing training efficiency using embedded pedagogical tools in a combat flight simulator. In *Proceedings of the Human Factors and Ergonomics Society 49th Annual Meeting* (Orlando, FL). Human Factors and Ergonomics Society: Santa Monica, CA.

TABLE OF CONTENTS

INTRODUCTION

Imagine being behind the stick and throttle of a military fighter jet. You're cruising at high altitude when you suddenly realize there is an enemy approaching you rapidly. You start to maneuver in order to avoid the threat, but the enemy anticipates your moves and suddenly you're engaged in a one-on-one dogfight. You maneuver your aircraft and try to become a threat to the attacker rather than the target. But the enemy is better than you and is slowly gaining advantage over you.

Is this for real, or are you in a training simulator? Does it matter if it is real or not? What are the differences between flying a real aircraft and in a simulator? Is your level of stress the same? Is your commitment to performing the job the same? In which situation would your learning be better?

Today, training simulators are widely used for a number of reasons. Economic reasons, safety, environmental concerns and accessibility (Angelborg-Thanderz, 1990; Jorna, 1993) are all strong arguments used to promote the use of simulators instead of training in the real environment. Simulators can provide opportunities of repeated practice in a safe and controlled environment (Kneebone, 2003). Using simulators for training has other advantages. An instructor can assess the behavior of the trainee and provide adequate feedback and debriefings. A simulator is often the most appropriate place to practice coping with emergency situations; often in a more realistic way than is possible to do safely in the real world. In military applications, simulators are used for exercises using weapons in scenarios that are impractical – if not impossible – to do in peace time. Simulators are also excellent for testing of new procedures and/or new equipment and to practice emergency procedures, as well as being a great tool for research. Flying aircraft in the real world requires coordination with numerous other services (such as air traffic control, maintenance), and need appropriate weather and visibility conditions (Lee, 2005).

In aviation training, simulators are assumed to provide good transfer-of-training. A simulators' training value is often assessed only through the degree of technical fidelity, for instance, degrees of visual field, latency times of the visual system, motion systems, etcetera (Borgvall, Castor, Nählinder, Oskarsson & Svensson, 2008). In some cases, a simulator with low technical fidelity can provide excellent training (Talleur, Taylor, Emanuel Rantanen & Bradshaw, 2003). The opposite is also true – even a very high fidelity top-of-the-line simulator might lack the possibility of getting the user involved to such a degree that meaningful training can be achieved (Nählinder, 2006). However, in order to assess a training simulators' fidelity from a Human Factors point-of-view, the user of the simulator must be considered (Salas, Bowers & Rhodenizer, 1998; Longridge, Bürki-Cohen, Go & Kendra, 2001).

Bell & Waag (1998) concluded that "Although a fair amount of opinion data exists that suggests there is training potential in using simulation, actual transfer data are extremely limited". Even though there is some increase in transfer-of-training studies (Taylor, Talleur,

Emanuel & Rantanen, 2005) they are still rare. This might be due to the fact that such studies are costly and cause logistical problems (Lee, 2005).

Simulators today have a much higher degree of fidelity than yesterday, but does this mean that learning is better? Does higher fidelity automatically lead to higher degree of transfer-of-training?

The benefit of motion cueing has been debated. Some studies show that motion cueing does not improve the effect of training (Bürki-Cohen, Booth, Soya, DiSario, Go & Longridge, 2000), while others have found a small positive effect (Vaden & Hall, 2005).

In a study on medical training, the purpose was to compare two simulators with respect to training effectiveness. The authors found no difference in learning between a high-fidelity full-scale anesthesia simulator and a computer screen-based simulation. They conclude that the choice of training device should be made "on the basis of cost and learning objectives rather than on the basis of technical or fidelity criteria" (Nyssen, Larbuisson, Janssens, Pedeville and Mayné, 2002).

Certain aspects of fidelity continues to increase, but flight simulators often completely lack fidelity in certain areas. The visual presentation, sound and motion cueing might be very realistic but radio communication (Bürki-Cohen, 2003) and realistic weather simulation often lack fidelity – if it is at all simulated (Perey, 2008)!

How can a simulators' potential to provide effective training be measured? What should be measured and how should the data be interpreted?

Purpose

The purpose of this thesis is to describe how the training potential of a flight simulator can be assessed. The thesis describes both what to measure, and also how to analyze and interpret the data. It also describes how the training potential of a simulator can be increased further.

This thesis is a product of applied Human Factors research and therefore there is an emphasis on how to deal with research in a very much applied, complex and dynamic world. There is a strong focus on the aviation context and more specifically on flight simulation training. However, the ideas are applicable to other areas as well. The thesis will contribute, not only from the results of the studies, but also from the experiences and thoughts on performing research in an applied, practically focused environment.

ASSESSING TRAINING POTENTIAL

To improve simulator training there is a need for increased knowledge about the transfer effects between simulator and reality. This in turn means that there is a need for increased knowledge about the mental workload experienced in respective environment. Relying on ratings alone is "major problem with research in this area" (Hays, Jacobs, Prince & Salas, 1998) and may mislead the attention of research to pursue ever increasing request for higher fidelity.

High fidelity does not guarantee high training value. Rather, it is important to focus on the human individual and his/her mental capacity. By collecting data allowing a comparison of mental workload and physiological reactions between a simulated and a real world, it is possible to understand more about the learning potential in these situations.

Mental workload

Mental workload is often used to evaluate system design, mission and training. Many studies have been interested in finding an unobtrusive, reliable and quick real time assessment of mental workload. In aviation, the environment is often dynamic, where sudden changes on the demands of human operators can be difficult to handle efficiently. According to Roscoe (1992), there is evidence that the failure to perceive the mental demands of a flight task correctly has been a causative factor in several aircraft accidents. Reliable, valid and dynamic measures of mental workload are therefore usable in a variety of situations (Castor, Hanson, Svensson, Nählinder, LeBlaye, MacLeod, Wright, Alfredson, Ågren, Berggren, Juppet, Hilborn, Ohlsson, 2003).

Mental workload is sometimes referred to as "the portion of the operator's limited capacity actually required to perform a particular task" (O'Donnel & Eggemeier, 1986), or as "the difference between the capacities of the information processing system that are required for task performance to satisfy performance expectations and the capacity available at any given time" (Gopher & Donchin, 1986), or "as the effort invested by the human operator into task performance" (Hart & Wickens, 1990). Mental workload is a complex concept and might be difficult to capture with a single measure (Magnusson, Berggren, Danielsson & Svensson, 2001).

Mental workload has long been used in the aviation domain for evaluating task efficiency, training efficiency, and evaluation of aircraft design as well as for mission analysis and assessment of pilot performance during flight operations (Svensson, Angelborg-Thanderz, Sjöberg & Olsson, 1997; Wickens & Hollands, 2000). Mental workload has been shown to be a precursor to operative performance mediated through the concept of situation awareness (Endsley, 1995; Alfredson, 2007), in a variety of studies (Nählinder, Berggren & Svensson, 2004), see Figure 1.

Figure 1. The relationship between the concepts of Mental Workload, Situation Awareness and Operative Performance.

Being a precursor for performance, mental workload is an important measure of success. However, performance might be biased and subject to variations due to circumstances beyond control.

The relationship between mental workload, situation awareness, and performance shows a pattern that is reoccurring in several studies, over different participants and in both real and simulated flight, in military and civil settings. There is a causal and logical relation connecting mental workload with situation awareness and situation awareness with performance. An increase in mental workload (a more demanding task) leads to a decrease in situation awareness, which in turn leads to lower performance (Nählinder et al., 2004). The results are also an empirical justification of the use of the concept situation awareness.

Measurements of mental workload

The most common way to measure mental workload is by using self ratings. This method has its limitations, since self-ratings can be biased and only can be measured at certain discrete points in time.

In applied research it is necessary to use measures that are easy to implement, are unobtrusive, and can be measured dynamically in real time (Backs & Boucsein, 2000; Wilson, 2002a). The participant should be able to perform the task as usual without being distracted by the equipment. Recording these variables in real aircraft – especially in military aircraft – further restricts the number of usable measures.

Psychophysiological measurements are of particular importance in the flight environment, since they offer a non-intrusive method to collect objective, dynamic data from pilots and other crewmembers (Wilson, 2002a; Wilson, 2002b). Psychophysiological data can be recorded continuously, thereby providing information on the crews' reactions in a highly dynamic environment. Once the equipment is in place, the recording can start and will run until the test is over without any intervention from the experimenter or the participant.

Several psychophysiological measures have been candidated for measuring mental workload: heart rate, heart rate variability, peripheral blood flow, oxygenation, eye movements, blinks, muscle activity/strain, skin temperature at various locations on the body, electrodermal activity, hormones, blood glucose, electromyography (EMG) (Hewson, McNair, & Marshall, 1999), brain activity (Hankins & Wilson, 1998; Sterman & Mann, 1995; Wilson, 1993) to mention a few (Boucsein & Backs, 2000).

Psychophysiological data is often found to be sensitive and a good complement to ratings in aviation research (Lee & Liu, 2003). Used in combination with ratings psychophysiological data can draw a reliable picture of the dynamics of mental workload and its effect on situation awareness and subsequently on operative performance (Nählinder et al., 2004). Advanced statistical techniques are used for combining psychophysiological, ratings and performance data into models of human behavior (Angelborg-Thanderz, 1990; Svensson & Wilson, 2002; Rencrantz, Lindoff, Svensson, Norlander & Berggren, 2006).

Heart rate can be measured at quite high temporal resolution, as a heart beat occurs about once a second. The changes of level of mental workload might not be as quick. It is likely that in applied situations mental workload varies only slightly from one minute to the next. Ratings of mental workload are normally not as often as that, but heart rate (as an indication of mental workload) can be measured much more often than that.

There are several psychophysiological measures that can be used in applied situations. Most commonly, activity of the *heart*, the *eye* and the *brain* are used to assess mental workload.

Heart

Heart rate has proven to be a sensitive measure that can be used to assess mental workload (Eggemeier, Biers, Wickens, Andre, Vreuls, Billman & Schueren, 1990). Heart rate is reliable (Angelborg-Thanderz, 1990; Svensson, Angelborg-Thanderz & Wilson, 1999) and proven to be quite sensitive in the aviation environment (Roscoe, 1993). It has often been used in combination with ratings (Roscoe, 1987; Roscoe 1992).

The hearts activity is fairly simple to measure. There are several unobtrusive easy-to-use pieces of equipment available. The downside of using heart rate as a measure of workload is that the heart is affected also by physical workload (Dahlström & Nählinder, 2006). As long as the pilot is seated during flight, stable and reliable heart rate can be measured without interference from any significant other muscular activity (Wilson, 2001).

The hearts activity is often measured in pulse (beats per minute). Since the pulse measure does not follow a normal distribution, inter-beat interval (IBI) is often used instead. This is of course an advantage when doing statistical testing. However, there are other ways to make sure that data is normally (or almost normally) distributed.

Heart rate data can also be used to calculate heart rate variability. Heart rate variability is a measure of the variation in the heart rate over a period of time. The hearts' activity is affected by a number of factors, such as physical activity, time of day, stress, oxygenation, metabolism, hormones, medical substances, etc. Even while resting the heart rate is not constant. Heart rate variability is a measure of this variation. Jorna (1993) discussed the usefulness of using heart rate and heart rate variability as a measure of mental workload in real flight. Heart rate variability was found to be quite crude, and could not distinguish between different levels of mental effort.

When calculating heart rate variability, it is very important that each heart beat is measured correctly. Failing to register a beat or registering a beat when there was none will seriously influence the heart rate variability measure (Berntson & Stowell, 1998).

Eye

Many languages have a saying – eyes are the mirror of the soul. This saying is tracked back to Cicero (106-43 B.C.), who is quoted as having said, *"Ut imago est animi voltus sic indices oculi"* (The face is a picture of the mind as the eyes are its interpreter). Eye movements are interesting to measure, since they reflect your visual information search, and eyes are believed to provide information about the mental state of a person. It is often easy to see if a person is happy, sad or angry by simply looking him/her in the eyes.

Higher fixation frequency may indicate higher workload (Svensson et al., 1999). Svensson et al. (1997) found that shorter fixation-times looking out of the cockpit and longer fixation times looking at instruments indicated higher information load in a military aircraft. Eye movement activity can be measured in several different ways, each way providing somewhat qualitatively different data (Alfredson & Nählinder, 2002; Alfredson, Nählinder & Castor, 2004).

Brain

EEG, electroencephalography, has also been shown to be useful as a measure of mental workload (Sterman & Mann, 1995; Wilson & Russell, 2003). EEG is measured by placing electrodes on the head of the participant. These electrodes pick up the very weak signals (μV, microvolt) that are recorded. Since the signals are weak, EEG is very sensitive to external noise such as muscle activity and electromagnetic fields (Nählinder, 2006). Because of this, EEG is difficult to measure in noisy environments.

Future of psychophysiological measurements

Faster, better, simpler, cheaper and less intrusive recording equipment is requested and would greatly facilitate research in the future. It is a great advantage to be able to perform measurements without having to stick electrodes on the body. Off-body sensors, wearable body sensors, camera based measuring and self calibrating equipment allow for unobtrusive measuring. However, it is also important to be able to handle and analyze the vast amounts of data generated by psychophysiology recording equipment.

Adaptive aiding

Psychophysiological data have also been used to automatically assess someone's mental status and to feed this information back to the system, which then knows if the person is in need of help or not. If so, the system can adapt automatically. This is called adaptive aiding (Rouse, 1988).

Adaptive aiding means that a system can adjust itself according to the current mental capacity of the operator. For instance, an aircraft can automatically notice that the pilot is becoming mentally overloaded to such an extent that his or her performance is becoming degraded. In such a state, the performance will be impaired and the operative outcome might become catastrophic. Interesting work is being performed in this area (Wilson & Russell, 2003; 2007), but more research is needed.

In an adaptive aiding environment, the system (aircraft) will adjust to relieve some of the stress on the operator. The long term goal is to adapt the information presented to the pilot in real-time as an integrated part of a decision support system (Alfredson & Nählinder, 2002).

METHODOLOGICAL CONSIDERATIONS

Often, differences between subjective and objective measures are stressed. However, objective measures are seldom, if ever, truly objective. For instance, heart rate is often claimed to be an objective measure. But heart rate is only an objective measure of the participants' heart rate. It is *not* an objective measure of anything else (such as mental workload). Further, the analysis and interpretation of heart rate data is highly subjective! On the other hand, a self-rating of mental workload is truly a measure of mental workload (therefore being objective), but might be biased, because of human aspects (therefore being subjective). For the reasons above, it should be clear that the division into objective and subjective measures is not meaningful! A much more interesting description of the quality of data is the data's reliability and validity.

Reliability of measures

Reliability is the degree to which a measure will give the same results if administered twice (or repeatedly) under the same circumstances (Heiman, 2001). A reliable measure will produce the same results each time, and is often described as precision. Reliability is inversely related to "random error" and does not automatically imply validity; a measure can be reliable without being valid. This occurs if the measure always is off the true value each time by the same amount. A speedometer in a car may, for instance, always show 115 km/h when the car is really travelling at 120 km/h. The speedometer is reliable (it measures consistently), but not valid (since it deviates from the true value by some percent).

Validity of measures

Validity is the degree of which a measure really measures what it is supposed to measure (Heiman, 2001). A valid measure will give a correct description of the object it measures. Validity is often described as accuracy. Validity does not automatically imply reliability. A measure can be valid without being reliable. This occurs if the measure varies around the true value. In the example above, this would mean that the speedometer would show varying speed (for instance 115, 125, 120, 110, 130 km/h) when the car is travelling at 120 km/h. In this case, the speedometer is not reliable (since it is not consistent), yet it is valid, since it measures (on average) the true speed of the car.

Improving the reliability and validity

Field studies improve the generality and validity of the results, but the data that is collected may be influenced by a number of factors over which the study does not have control (noise). In order to create valid measures in a noisy environment, several similar variables can be combined. By amalgamating similar measures into a single latent variable or factor, this new is a more reliable and more valid measure than each of the single measures. The factor is likely to be more valid, since the signal to noise-ratio is increased by increasing the number of signals (variables). It also becomes more reliable since random error is more likely to be canceled out. Therefore, it is a good idea to measure the same phenomenon using different

measures (variables), and combine these measures to a single, valid value. This is called construct validity (Heiman, 2001). A single question in a questionnaire may not be perfect (reliable and valid), but combining several questions into one factor may produce a measure (factor), which is both reliable and valid.

For instance, when assessing "mental workload", it is a good idea to let the participant perform self-ratings of mental workload as well as closely related concepts (such as "stress", "mental spare capacity", etc). By simultaneously gathering instructors or peer students ratings of the participants' mental workload, one can further increase the quality of the data. These different variables can be combined using statistical methods to create a single valid value of mental workload. The use of expert ratings, instructor ratings, psychophysiological measures, behavior aspects etcetera might further increase the reliability and validity.

Capturing the dynamics

Variations in a single measure during a continuous trial are often sought to be assessed. Measuring variations require dynamic measures. There are several ways in which a measure can be assessed repeatedly during the course of the trial.

A *discrete* measure is a measure which is only measured a couple of times during a scenario, for instance, a questionnaire that is administered before and after a mission (or several missions).

An *event related* measure is taken before, during or after a certain event, for instance, each time a pilot performs a touch-and-go procedure. This approach is useful mostly if the occurrence of each event of interest is known in advance.

A measure can also be taken at certain time intervals; this is sometimes called *quasi-dynamic* (Alfredson, Angelborg-Thanderz, van Avermaete, Bohnen, Farkin, Ohlsson, Svensson & Zon, 1997). For instance, a participant rates the level of mental workload every five minutes. This is useful if a long duration procedure is to be followed or when the exact starting time of an interesting event is not known until after the trial. Quasi-dynamic measures are also useful for capturing variations during the day such as sleepiness, fatigue, arousal, etc.

A *dynamic* measure is measured continuously, for instance, a psychophysiological recording that keeps measuring continuously regardless of the evolvement of the scenario.

Dealing with individual differences

People are people. The heart activity, for instance, is quite different between different people. Some individuals have a high resting heart rate, whereas others have a lower. Some people are more responsive than others, that is, their heart rate will increase more when presented with a certain stimulus. Further, a person might be different from one day to another. Individual differences may even be larger than the differences within each individual, which will make statistical tests fail to identify interesting effects (Magnusson, 2002; Dahlström &

Nählinder, 2009). The differences between individuals do not only apply to psychophysiological measures, but also to ratings. Some individuals tend to rate aspects at the extreme ends of the scale, while others keep their rating more along the middle of the scale.

One way of reducing the variation in data is to use statistical methods that only focus on the differences within participants. Let each participant perform the same task under different conditions and then analyze the difference between the conditions for each participant. This is called a within-participant design (Keppel, 1991). In this type of design of an experiment, each participant is exposed to all levels of the independent variable. This design will remove individual differences regarding level, but the differences in responsiveness can still make statistical methods fail to find interesting results. Data can be adjusted for a base-line measurement, such as a resting value. Individual responsiveness can be adjusted for by using standard provocations assessed together with the baseline registration.

There are mathematical methods available to eliminate individual differences, facilitating the possibility to find the differences between experimental conditions. The methods below completely erase the differences *between* individuals, so they can only be used when using a within-participant design. The methods suggested below do not require base-line measurements or standard provocations, and have been used in several of the studies in this thesis in order to eliminate individual differences. Standardizing data and/or normalizing data may reduce the degrees of freedom by two or three, respectively.

Standardize data

To reduce the individual differences, one can choose to standardize the data. The main purpose of standardizing data is to make the data similar and comparable between individuals (Magnusson, 2002). Standardized data removes the individual differences. Thereby it enhances relative changes for individuals, so within participant effects can more easily be spotted. Standardized data always get the mean value zero, and the standard deviation one, see Figure 2.

In the figure below, x_1 is the new value to replace the old x_0. $\bar{x}$ is the mean value and s is the standard deviation for this participant's values.

$$x_1 = \frac{x_0 - \bar{x}}{s}$$

Figure 2. Standardize data.

Standardized data is comparable between two individuals, but since the average is zero and standard deviation is one. As such, standardized data this is much less intuitive than the original data. For instance, if someone had a heart rate of 60 beat per minute which later increased to 100, the standardized value might say that the heart rate was .8 and increased to 1.3, which obviously is less intuitive.

The suggestion is to restore values. The restored data will be similar to the original values, but all individuals will have the same average and same standard deviation. From a statistical point of view, this method is no different than the method above, but the data is easier to relate to.

$$x_1 = \frac{x_0 - \bar{x}}{s}\sigma + \mu$$

Figure 3. Standardize and restore data.

The formula in Figure 3 is very similar to the formula in Figure 2 with the exception that the data in Figure 3 has been moved to match the original data's average value and standard deviation. σ is the standard deviation for all participants, and μ is the average of all participants (grand mean).

This method has the same statistical properties as standardizing the data, but it is easier to interpret, since the values are restored. The values are *not* restored to the individuals original values (this would be pointless, of course). Rather each individual will have the same average value as the average value of all the individuals' original values. The same relationship applies to the standard deviation: each individual's standard deviation will be the same as the average standard deviation.

The advantage of the proposed standardization is that reactions to a stimulus will be more apparent, since inter-person variation is reduced. The disadvantage is just that: the difference between individuals is completely erased.

Figure 4 shows an example of standardization of data. The left side shows the raw data (unstandardized) and the right side shows the same data after it has been standardized. Note that individual differences are removed, but the common increase that occurs just before the middle becomes much more apparent. The data will remain at approximately the same values. Before the sudden increase, values lay around 15 and at the increase around 28.

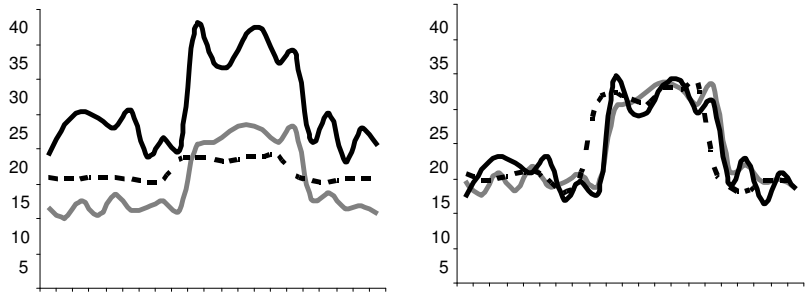

Figure 4. An example of standardization of data. Unstandardized data (left) and the same set of data after being standardized (right).

Normalizing data

Many statistical methods require that the data follow a normal distribution (at least to a certain degree). However, many times, the data does not. This is especially true with psychophysiological data, and to some degree also with rated data. One solution is to normalize the data, making it follow a normal distribution (Nählinder, 2006; Dahlström & Nählinder, 2009). Each value of a dataset is replaced with a score corresponding to its estimated cumulative proportion in the distribution, effectively transforming virtually any distribution to a normal distribution (or at least as close as possible).

The disadvantage of normalizing data is that it cannot be undone. The scale of the measure becomes arbitrary. After normalizing heart rate data for instance, the actual heart rate values are no longer known. In the normalized dataset, the average value will be zero, and the standard deviation will be one.

Conclusions

From a research perspective, the conditions in which studies are performed can be suboptimal in many ways. Performing studies in real-world settings may cause the collected data to be noisy and contain un-wanted variations, concealing the effects of interest. The statistics will be weak and it will be difficult to draw valid and reliable conclusions from the results.

However, there are techniques available to improve the quality of the data. Reliability and validity can be improved by carefully selecting what measures to collect. In many cases, dynamic measures are required to capture and explain a situation as it evolves. Further, individual differences can be eliminated. This allows for easier detection of within-individual effects, but will completely remove between-individual differences. Finally, normalizing data allows for stronger statistical testing of non-normally distributed data.

THE STUDIES

In several of the studies in this thesis, psychophysiological assessment of pilot mental workload has been examined. The relationship between psychophysiological assessment of mental workload and ratings of mental workload has been analyzed. In two of the studies, similarities and differences in psychophysiological reactions between simulated and real flight were analyzed.

All of the studies focus on pilot's singlehandedly flying an aircraft (real or simulated). The participants in the studies are professionals. Their level of commitment to their work is high and they make an effort to perform well. The amount of mental workload they experience ranges from high to very high. In the studies, the pilots were very much focused on the task and on performing as well as possible.

Study I: Psychophysiological reactions in air-to-ground missions

A study (Magnusson, 2002) was performed at the air force wing F17 in Kallinge, Sweden. The purpose of this study was to measure similarities and differences in psychophysiological reactions and ratings between simulated and real flight. Five male fighter pilots participated in the study. They flew the exact same air-to-ground mission three times in a simulator and later three times in real flight. In both cases, the same type of scenario, the same tactics and the same type of aircraft was used.

The participants flew an air-to-ground mission which has been divided into four phases. First, the pilots flew towards the target area. In the second phase (the pre-attack) they were required to fly at high speed and at low altitude. Next, in phase three (the attack), they attacked a ground target by performing a pop-up maneuver and delivering the weapons. Finally, in phase four (disengagement), they disengaged and flew back home. This air-to-ground mission was carefully selected in order to make it as similar as possible in the simulator and in real flight.

The aircraft was a JA37 "Jaktviggen". This aircraft was the fighter aircraft in service at the air force wing F17 in Kallinge in 2002, however today, the JA37 is no longer in service in the Swedish Armed Forces; it has now been replaced by the JAS39 Gripen. The JA37 simulator had a realistic cockpit environment (built from a discarded aircraft) and good out-of-the-window view of the surrounding virtual environment. Even though the simulator has a motion cueing capability, this was switched off during the study. Thus, the simulator had no motion.

The participants' heart rate, heart rate variability and eye movements were measured using a VITAPORT digital portable recorder (VITAPORT II digital recording device from Temec Instruments BV, Gemert, the Netherlands). The VITAPORT is a small (40 x 90 x 150 mm), lightweight (750 grams, batteries included), portable digital recorder (Fahrenberg & Wientjes, 2000). The participants also rated their mental workload, situation awareness and performance both in the simulator and in the real flight. The psychophysiological data was standardized in order to focus on the similarities between the participants, removing their individual differences.

Results

The results from the psychophysiological reactions clearly show that there was no difference in the way the participants reacted in the simulator compared to how they reacted in real flight. At the time of weapons delivery, there was a distinct increase in heart rate, decrease in heart rate variability and decrease of eye movements, in the simulator as much as in real flight, see Figure 5. The correlation in psychophysiological reaction between the simulated and real flight was very high, especially for the heart rate and heart rate variability measures. The eye movements were also highly correlated, but not quite as high.

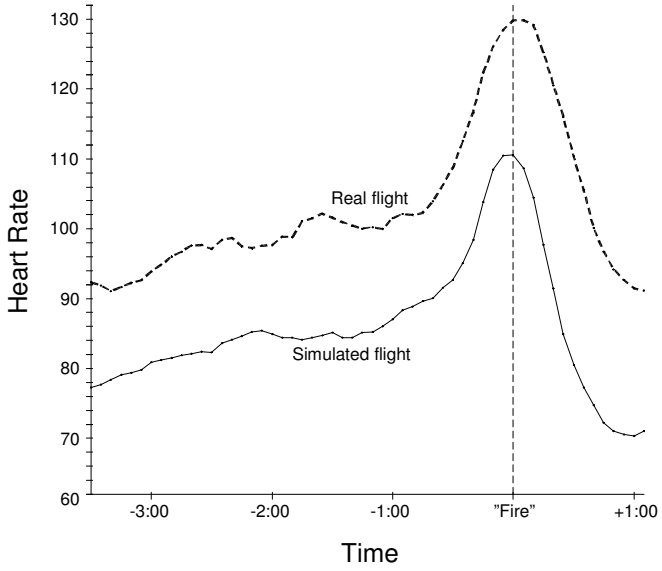

Figure 5. Similarities and differences in psychophysiological reaction. The data has been standardized.

There was a difference of level. That is, the heart rate was higher when flying the real aircraft than it was in the simulator, but the reactions (the pattern) were very similar.

The results also show a difference over the three times (sorties) the mission was flown. The first sortie, the heart rate was higher than the two other sorties. This was true both for the simulated and real flights (see Figure 6). The statistical analyses show a significant main effect of Time $F(83, 913)=13.8$; $p<.001$. There is also a main effect of Sortie $F(2, 22)=11.63$; $p<.001$ and a significant interaction between Time and Sortie: $F(166, 1826)=1.34$; $p<.005$. Post-hoc analysis indicates that there is a difference between the first sortie and the other two.

There is a main effect of Type of flight (simulated or real), see above, but there is no interaction between Type of flight and Time (or Sortie). This means there was no difference in reaction between simulated and real flight (or between sorties), even though there is a difference of level. The curves below are (statistically) parallel.

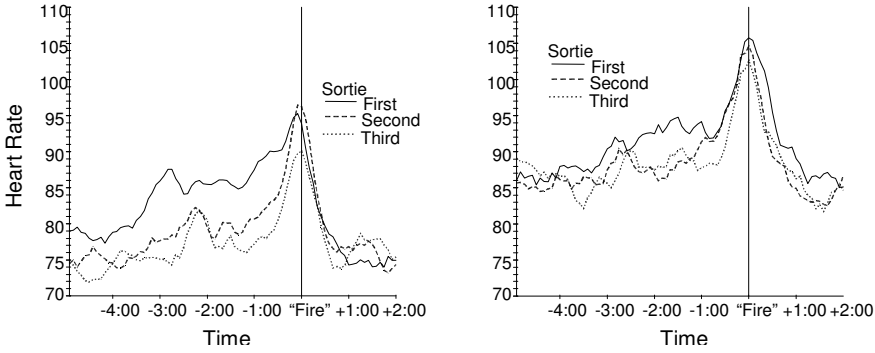

Figure 6. The heart rate during three consecutive flights. Simulated flight to the left, real flight to the right. The data has been standardized.

For heart rate variability, the same story is true, see Figure 7. The first sortie, the heart rate variability was lower than the two other sorties. Again, this was true both for the simulated and real flights. The statistical analysis show a significant main effect of Time $F(83, 913)=11.89$; $p<.001$, as well as a main effect of Sortie $F(2, 22)=11.99$; $p<.001$. There is also a significant interaction between Time and Sortie $F(166, 1826)=1.23$; $p<.05$. A post-hoc test indicates a difference between the first sortie and the other two.

There is a main effect of Type of flight (simulated or real) as shown above, but again, no interaction between Type of flight and Time (or Sortie). This means that even though there is a difference in level between simulated and real flight (and between sorties), the curves are parallel, and the pilots' reactions are indeed very similar.

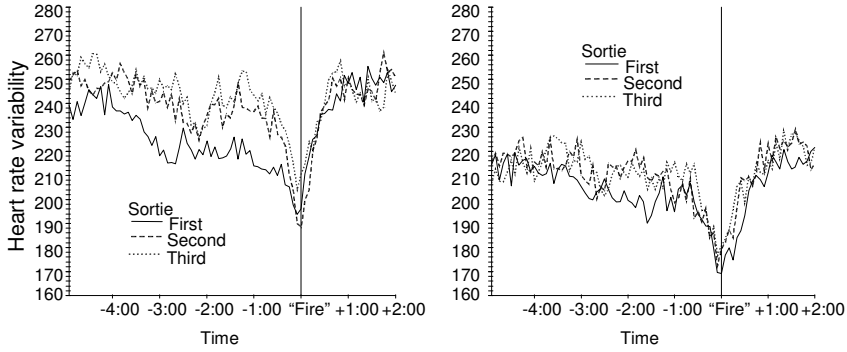

Figure 7. Heart rate variability during three consecutive flights in simulator. Simulated flight to the left, real flight to the right. The vertical scale is arbitrary. The data has been standardized.

Eye movements also show that there is a high degree of similarity between simulated and real flight, see Figure 8.

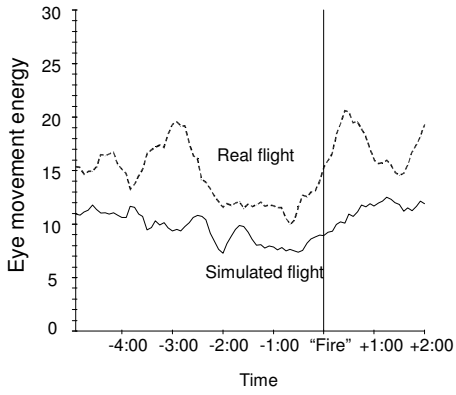

Figure 8. Eye movement energy for simulated and real flight. The vertical scale is arbitrary. The data has been standardized.

In the study, there was a high degree of similarity in the psychophysiological reactions between the simulated and real flight. There was also a decrease in heart rate, increase in heart rate variability and decrease in eye movements from the first to the second and third repeated sorties, both in simulator and in real flight.

The study (Magnusson, 2002), was published in International Journal of Aviation Psychology.

Study II: Mental workload and psychophysiological reactions

In study II, Magnusson & Berggren showed that the participants' ratings from the simulated part of the study above (Study I), show a very similar pattern to the psychophysiological data (Magnusson & Berggren, 2002).

Results

At the time of the attack (when the weapons were fired), the participants rated the highest level of mental workload, which is when they had the highest heart rate, see Figure 9. The correlation between the two measures was fairly high - .56, $p<.01$.

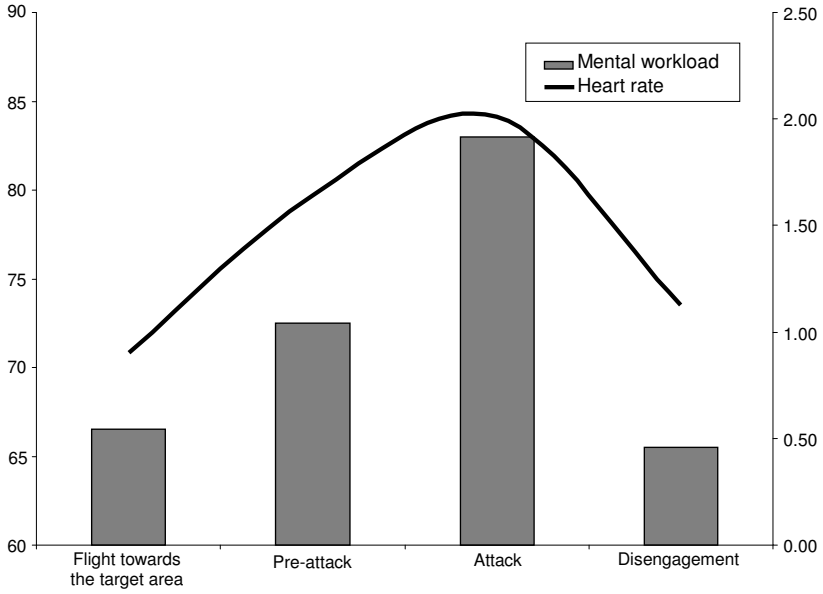

Figure 9. The similarities between heart rate (scale on left side) and ratings of mental workload (scale on right side). The data has been standardized.

The participants' ratings can (together with heart rate) be described in a LISREL model (Jöreskog & Sörbom, 1984) showing the relationship between mental workload, heart rate, situation awareness and performance, see Figure 10. The model shows that, as mental workload increases, the heart rate increases and the situation awareness decreases. As a result, performance decreases as well.

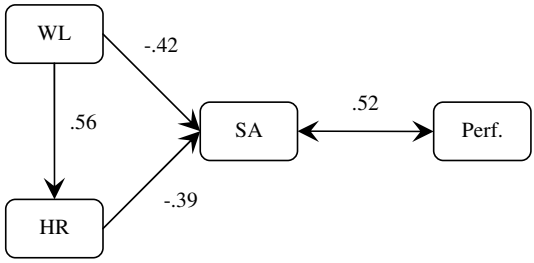

Figure 10. A LISREL model describing the relationship between Mental Workload (WL), Heart rate (HR), situation awareness (SA) and performance (Perf). The model is statistically significant.

The second study (Magnusson & Berggren, 2002) was presented as a poster at the Human Factors and Ergonomics Society Annual Conference (HFES) which was in Baltimore, MD, USA. The paper was written by the two authors equally. Magnusson performed the analyses and presented the poster at the conference.

Study III: Comparing two recorders

The purpose of this study (Dahlström & Nählinder, 2006) was to investigate the possibility of using a certain heart rate-recording equipment that is simpler to use, cheaper and less intrusive than the VITAPORT equipment used in study I and II. The study was performed at the Lund University School of Aviation.

Heart rate data recorded by cheaper consumer-type psychophysiological equipment (Polar Team System) was compared to data recorded by the more expensive VITAPORT recorder. The Polar Team System is easy to use, non-intrusive and a large number of recorders can be employed simultaneously at low cost.

In the study, student pilots flew a profile in the simulator which was part of their syllabus. This profile was later flown in real flight. In both the simulator and in real flight, the students were equipped with both heart rate-recording equipments: the VITAPORT and the Polar Team System.

Great effort was invested into synchronizing the equipment and to make the comparison as fair as possible. Since the purpose of this analysis was to compare the raw data from the two systems, data were not standardized and not normalized.

Results

Analysis show that the recorded heart rate was very similar in the two pieces of equipment, see Figure 11. The correlation between the two systems was .981, indicating that 96.3% of the variation in the Polar Team System heart rate data can be explained by the VITAPORT heart rate data.

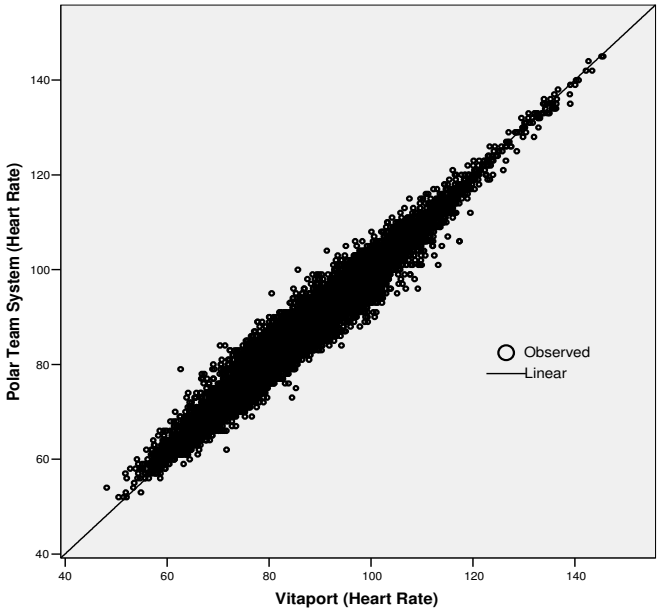

Figure 11. A scatterplot of the heart rate data for the two systems.

The remaining 3.7% of the variation between the two systems can partly be explained by the following issues:

1. The Polar heart rate data are always rounded off to a whole value (beats per minute), whereas the VITAPORT data are calculated with high precision.

2. Internal rounding off in the Polar recorder may not have been accounted for in this analysis.

3. Even though the two systems were synchronized initially, there is no guarantee that the systems record the same five-second-intervals.

4. The internal clocks of the two recorders may differ after a day or so of measurements.

5. The two systems might handle arrhythmias and unusual heart activity and artifacts differently.

If controlled for, these issues will most likely further decrease the differences between the two systems.

The results show that the Polar Team System is an excellent alternative to the VITAPORT when it comes to recording heart rate data. It produces very similar results to the VITAPORT recorder. However, the VITAPORT has higher resolution and have the possibility of

calculating other variables, such as heart rate variability, whereas the waist-band equipment is more limited.

The high degree of correlation between in-flight heart rate data collected by the VITAPORT and Polar Team System recorders, shows that the less intrusive and less costly equipment is a reliable and a cost effective alternative to the clinical and research oriented device. This is important because it allows easier collection of heart rate data. It allows data to be collected routinely, for instance at a flight school.

This study was published in Applied Psychophysiology and Biofeedback (Dahlström & Nählinder, 2006). The authors are listed in alphabetical order. The authors formulated the design of the study together. They were equally active in the data collection. Dahlström was responsible for scheduling and arrangements at the flight school and Nählinder performed the analyses.

Study IV: Psychophysiological reactions in a civil flight school

The purpose of this study was to investigate the similarities and differences between simulated and real flight (Dahlström & Nählinder, 2009) at the Lund University School of Aviation.

Eleven student pilots flew a profile in the simulator which was part of their syllabus. The profile consisted of several events, such as aborted take-off, take-off, engine failure, cruise, several repeated instrument approaches and finally landing. After successfully completing the simulator profile, the students flew a profile, which was very similar to the one in the simulator, in the real aircraft. The real flight consisted of the same events in the same order as they appeared in the simulator.

Heart rate, eye movements and ratings from eleven student pilots were collected. On all sorties (both in simulator and real flight), ratings were given by the participant, the instructor and an observer. The psychophysiological data was standardized and normalized in order to focus on the similarities between the participants, removing their individual differences.

Results

The results show that there is a high degree of correspondence between simulated and real flight for the various phases and events (see Figure 12 for heart rate and Figure 13 for self-ratings).

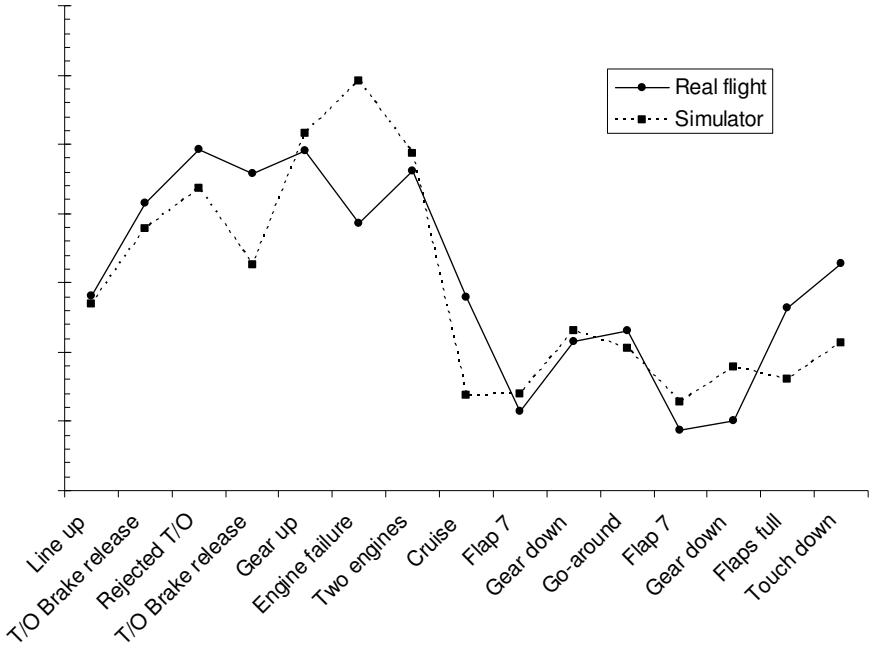

Figure 12. Heart rate data for various flight phases in simulated and real flight. The data has been standardized and normalized; therefore the vertical scale is arbitrary.

In this study clear similarities between simulated and real flight were found, much like in Study I. However, there were also some unexpected findings:

1. The rejected take-off caused differences in heart rate.
2. Heart rate during engine failure was higher in the simulator than in real flight.
3. There seems to be a mental (psychophysiological) preparedness for certain events.
4. The simulator landing produced no increase in heart rate.

There was a huge difference in heart rate reactions during the rejected takeoff between the simulator and the real flight.

The heart rates were higher during the engine failure in the simulator than it was in the real flight. This is the opposite of what would be expected: An engine failure in real flight should be more stressful than an engine failure in simulated flight. The engine failure in real flight occurred at safe altitude (with larger safety margins), whereas in the simulator, the engine failure occurred immediately after takeoff (at low altitude, with much small safety margins). Also, in the simulator, the engine failure was a “real” engine failure (initiated by an instructor at a control station outside the simulator), and the participants were required to perform the appropriate checklists in order to get the engine running again. In real flight, on the other

hand, the engine failure was simulated by the instructor putting the engine into idle, while telling the pilot that this is an engine failure. Therefore, the student pilots had to focus more and work harder in the simulator (causing the higher heart rate) than in real flight. It raises a philosophical question: which engine failure is most realistic - a simulated engine failure in real flight or a real engine failure in simulated flight?

There was an increase of heart rate in advance of an event in the simulator to larger degree than in real flight. The simulator engine failure occurred immediately after the Take-off and Gear up segments and it seems this was anticipated by the student pilots. They are aware of what moments will be practiced according to the syllabus, so it should be no surprise that they were aware of what was coming and they had time to be mentally prepared.

The landing phase was not at all realistic in the simulator; it has no visual system for out-of-the-window view. In real flight, the landings were visual (after an instrument approach). A landing in that simulator is simply done by making a controlled descent to zero altitude. There is no point in aligning the aircraft with a runway, since there are no runways. In a real aircraft, on the other hand, the landing is normally the most cognitive demanding flight segment (Wilson, 2001, 2002a).

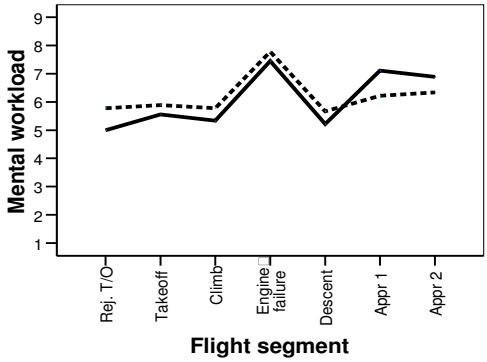

Figure 13. Student self-ratings of mental workload during seven of the flight segments. The solid line represents real flight and the dashed line represents simulator flight. Ratings were given on a scale from one to nine, where one is low and nine is high.

For the subjective ratings of mental workload (Figure 13), there was a main effect of flight Segment ($F[6, 48] = 9.22$; $p<.0005$). There was no main effect of Type of flight (simulated or real) and no interaction effect. Accordingly, this indicates that there was no difference in how the two types of flights were rated. The Engine failure produced the highest ratings of mental workload for both types of flight.

The forth study (Dahlström & Nählinder, 2009) is an article that is accepted for publication in International Journal of Aviation Psychology. Again, the authors worked together and are listed in alphabetical order. Dahlström was mainly responsible for the writing and Nählinder

performed the analyses of the psychophysiological data as well as of the ratings and questionnaires.

Study V: Psychophysiological reactions during aerobatics

In a study performed at the Lund School of Aviation in 2005 (Dahlström, Nählinder, Wilson & Svensson, 2009), various psychophysiological measures were collected during aerobatic maneuvering in real flight. Five instructor pilots flew an advanced aerobatics profile (in real flight only). The purpose of this study was to assess the usability of certain psychophysiological measures. The pilots performed a sequence of aerobatic maneuvers, such as roll, loop, "cuban eight", and "hammerhead". This sequence was repeated twice with a short straight and level flight in between. The participants' heart rate, eye movements and brain activity were measured. After the flight, the pilots rated mental workload, performance and difficulty for each of the maneuvers.

While heart rate is indeed a very easy and practical way to assess mental workload – even in real flight – it has several disadvantages. For instance, heart rate is very sensitive to physical workload. When performing aerobatic maneuvering heart rate is less likely to be sensitive to differences in mental workload since the physical load while have a great impact on the hearts' activity. Therefore, other psychophysiological measures could be useful. Besides heart rate (ECG), eye movements (EOG) and electroencephalogram (EEG) was measured as well as self-ratings. EEG has often successfully been used to discriminate between levels of mental workload (Wilson & Russell, 2003).

Results

Heart rate (Figure 14) and self-ratings (Figure 15) showed that the aerobatic sequences had the highest levels of mental workload. Heart rate alone could identify sequences of higher mental workload (during phases of low physical workload), while blink rate and eye movements were not able to identify phases of high mental workload. EEG data was difficult to analyze mainly due to large influences of muscle artifacts and problems with the recording of the data.

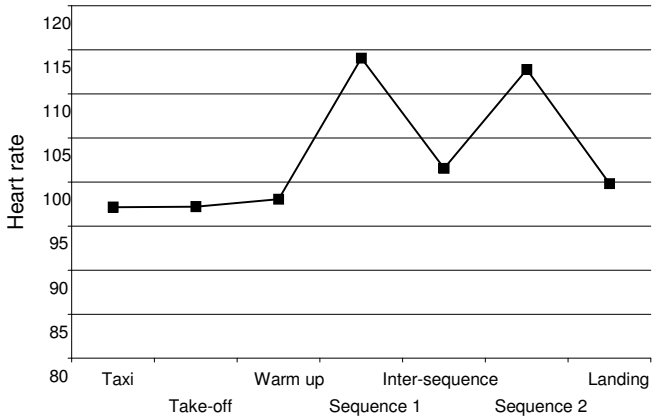

Figure 14. Heart rate during different flight segments. The data has been standardized.

The analysis shows a main effect of flight Segments. Post hoc analyses show that the ratings of mental workload were higher for the aerobatics sequences than for other flight Segments ($F[6, 30]=3.55$; $p<.01$), with landing being the highest thereafter. Ratings of Performance and Difficulty display a similar pattern ($F[6, 30]=3.27$; $p<.05$ and $F[6, 30]=5.18$; $p<.001$, respectively), with relatively lower performance ratings and higher difficulty ratings for the aerobatics sequences.

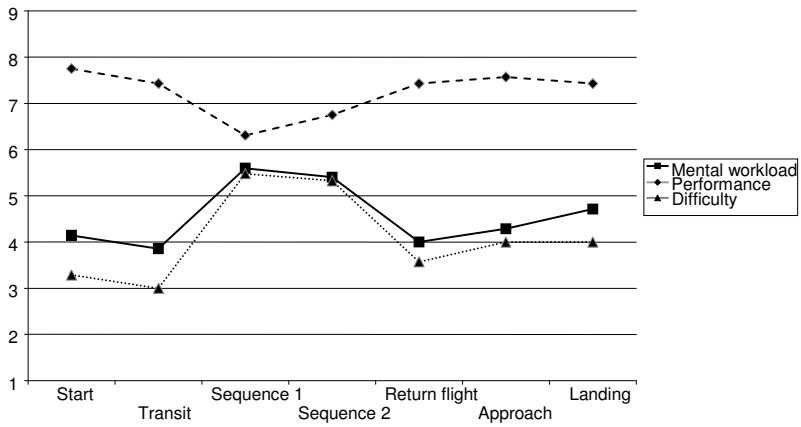

Figure 15. Average self-ratings of mental workload, performance and difficulty for the participants. Ratings were performed on a scale from one to nine where one was lowest and nine highest.

The study (Dahlström, et al., 2009) has been accepted for publication in International Journal of Aviation Psychology. Dahlström was responsible for the participants and the scheduling of the study. Nählinder and Wilson both performed analyses of the heart rate data, and Wilson

analyzed the EEG data as well as the eye blinks. The questionnaires were analyzed by Nählinder and Svensson.

Study VI: Embedded training tools

The purpose of the study (Nählinder, Berggren & Persson, 2005) was to evaluate the usefulness of a set of embedded training tools in a flight simulator, both from a student pilot and an instructor's perspective.

20 pilots from the Swedish Air Force Flying Training School participated in the study. Eight of the pilots were students and 12 pilots were instructors at the school. The students were all currently in basic tactical training ("phase 3+") of their flight education.

A flight simulator called "ACES" was used (Nählinder, 2004). It is a fairly simple flight simulator at the Swedish Defence Research Agency. It has two aircraft cabins (resembling SAAB JA37 "Viggen") and one instructor station. The instruments of the aircraft are displayed on a computer screen inside the cabin ("head down" display) and the outside the cabin world is displayed on a head-mounted display generating a stereoscopic view in excess of 100º field-of-view horizontally. The flight controls in the cabins (the stick, the thrust handle, pedals etc.) are real controls from discarded aircraft. The simulator is heavily focused on training of Within-Visual-Range ("dogfight") combat. The most important feature of the ACES simulator is its embedded training tools. These tools can help to illustrate and explain abstract relationships and parameters that are important to successful dogfight maneuvering. They have been implemented into the simulator to facilitate teaching and learning.

The participants rated the perceived usefulness of seven embedded training tools. Answers were given on a seven point scale, ranging from "little" to "much", or "easy" to "hard". For each of the seven training tools, the participants rated to what degree the following dimensions could be improved: *skill*, *knowledge*, *learning situation*, *understanding*, *usefulness* and *ability*. In all, each participant gave 7 times 6 = 42 ratings. For example, one question was "to what degree can *mistake correction* improve your *understanding* of dog-fight"?

Results

The ratings for each pedagogical tool are presented in Table 1, below. In this table, the ratings have been collapsed over the six dimensions. Students and instructors are not analyzed separately at this time. The seven training tools were all rated relatively high (see Table 1). This suggests that the participants strongly believe in the usefulness and suitability of such embedded training tools in a simulator training environment.

Table 1. Subjective ratings on seven different pedagogical tools.

	Mistake correc.	Changed pos.	Changed speed	Energy analyses	Camera view	Back-seat	3D traces
Average	6.03	5.46	3.76	4.83	6.03	4.90	5.84
Standard deviation	0.78	0.90	1.55	1.36	0.81	1.42	1.02

In this study (Nählinder et al., 2005), the design and preparation of the study was performed by Nählinder and Berggren. Persson (working at the Flying Training School) was responsible for the participants and the integration of the study into the scheduling and in accordance with the pedagogic approach of the school. Analyses of the data and writing of the study was mainly performed by Nählinder. The paper was presented by Berggren at the Human Factors and Ergonomics Society Annual Conference in Orlando, FL.

SUMMARY OF RESULTS

In the first study, similarities and differences in psychophysiological reactions between simulated and real flight were analyzed. The similarities are prominent: an increase in heart rate in the simulator is equally large as in real flight for the same flight phase. This is consistent between all psychophysiological measures. There is however, a big difference between simulated and real flight. The heart rate is constantly lower in the simulator. The heart rate variability is higher and the eye movements are less in the simulator.

In study II, the relationship between psychophysiological data and self ratings of mental workload was analyzed. The variations in these measures make it possible to combine the data into a statistical causal model. Heart rate is closely related to ratings of mental workload.

Two different psychophysiological recorders were compared in study III. The Polar Team System and the VITAPORT recorded simultaneously data on the participants. The two equipments were indeed very similar and produced almost exactly the same data. The VITAPORT has greater potential, but as tested in this study, they are equal in performance.

Similarities and differences in psychophysiological data were compared between simulated and real flight in study IV. In this study, there were no consistent differences between simulated and real flight as in study I. Heart rate did however differ on certain flight phases. The engine failure caused higher levels of mental workload in the simulator than it did in real flight. Rejected take-off and landing caused higher levels in real flight than in the simulator.

In study V, psychophysiological reactions during aerobatic maneuvering were studied. Heart rate and ratings of mental workload succeeded to correlate (at least for the low-g segments), while EEG and eye blinks failed to produce reliable results.

Study VI tested a concept of embedded training tools in a flight simulator. Students and instructors rated perceived usefulness of several such training tools. The results clearly indicate that embedded training tools can be of great value, both for students who are trying to learn new concepts, and for instructors who try to teach them.

CONCLUSIONS

Studying mental workload can be one way to assess training potential of a flight simulator. It focuses on the user, the trainee, rather than on the simulators' technical properties, such as fidelity.

Higher mental workload does not necessarily lead to higher training effectiveness. That is, the situation which has the highest workload is not requested. Training under high workload situations can be devastating from a learning perspective. Rather, the differences between simulated and real flight can be used to highlight discrepancies, which may be the subject of more detailed analysis.

In study I and IV, the pilots react similarly to a certain events regardless if they are performing the event in a simulator or in an aircraft. That is, the simulated flights seem to reproduce a similar experience to that of flying real aircraft. When there is an increase of heart rate in real flight, there is one in the simulator as well. Therefore, it is believed that training can occur in these situations.

The studies show deviations between simulator and real flight. These deviations or differences also provide valuable information about the relationship between simulated and real flight. For instance, in some cases the pilots experienced the simulator as having lower workload (assessed by psychophysiological measures) than real flight, and in other cases, the pilots experienced the opposite. Such points of deviations might be candidates for further in-depth analyses of learning possibilities in either environment.

The first study shows that the heart rate decreases from the first to the following two repetitions. This might be explained by the pilots gaining more experience and thus becoming more comfortable and more relaxed with performing the task. However, the pilots flew the mission first in the simulator three times (with decreasing heart rate), and after that in real flight three times (also with decreasing heart rate). However, there was an *increase* in heart rate from the simulated flights to the real flights, even though the pilots had gained experience in the simulator and therefore should have been more relaxed flying the mission the fourth, fifth and sixth time (in real flight). As noted previously, this is likely to be an effect of the situation being more dangerous, more engaging and more immersive than the simulated flights. Perhaps the participants would have had an even higher heart rate in real flight if they had not received training in the simulator?

Psychophysiological data can be easy to measure, study III showed that heart rate data can reliably be recorded by fairly simple and relatively inexpensive equipment. However, study V showed that EEG can be difficult to use in applied studies, being very sensitive to external noise.

The psychophysiological measures are unobtrusive, reliable and dynamic and provide data with high temporal resolution. Psychophysiological data in combination with individual and

instructor (and/or peer) ratings can be used to create reliable and valid measures of mental workload.

The results are very interesting and useful showing that psychophysiological reactions are indeed a good way of identifying similarities and differences between simulated and real flight. The results from the studies are believed to be equally valid in many other areas besides flight simulators.

Challenges for Human Factors methods

Today, applied research is performed in real world settings, and in close-to-real world settings such as simulators, mock-ups and in real environments interacting with simulated entities. Performing research in these environments generate data and knowledge that have higher representativeness and better generalizability than a laboratory study. However, from a research perspective, this is a mixed blessing.

The scientific methods available today are developed with laboratory studies in mind, using many participants, great control of independent factors, preferably with a classical experimental design and a dozen or two participants.

In applied research there are many challenges. There might be only a very limited amount of available participants. There is less control over independent factors. It might not be possible to perform a classical experimental-control group design. Running the trials is often costly and time-consuming and requires planning and scheduling. The amount of data available is often close to unlimited, with software loggings of hundreds of parameters.

The Human Factors research of tomorrow might be performed more “in the wild” (Hutchins, 1995), in applied settings outside the laboratory. New tools, new designs and new statistical methods are required to meet the demands of tomorrow’s human-factors research. These methods must be sensitive in finding results in noisy environments, using few participants and perhaps with little or no knowledge about baseline values.

The methods suggested in this thesis, such as eliminating between-participant variations and normalizing data to better fit statistical requirements, are small steps to improving the research methods. More work in this area is definitely needed!

Future of flight simulator training

The dogfight example in the very beginning of this thesis is one case when a simulator can provide a learning experience that might go beyond what could be learnt in a real aircraft. When training dogfight skills in a real aircraft, chances are your “enemy” is really a college acting as an enemy for you to train on. Which environment produces the most effective learning: fighting a real enemy in the simulator, or a simulated enemy in the real aircraft?

Simulators are often made to replicate the nature to as high degree of fidelity as technology allows. This is all very well, but as Human Factors scientists, we must question the higher

purpose (Hancock & Diaz, 2002). If the purpose of the simulator is to produce training, why does it look and feel as the real world? Is that the best it can do to teach us? Is it not possible that we might learn more from a simulator that is built to teach rather than to replicate?

Most simulators used for training are not in themselves good at teaching. Modern flight simulators lack the possibility of providing pedagogical feedback. The simulator may provide a very realistic environment in which training can take place, but the simulator itself is not a teaching device. It replicates a real world situation without providing pedagogical support that might facilitate learning. The results from study VI suggest that embedded training tools may have a great impact on training efficiency. There is a consistent opinion in favor of using the embedded training tools, implying that such tools may be the way forward in training simulator development. The study is a first step towards finding an optimal set of training tools that could be implemented in a training simulator facility.

Combining ideas from adaptive aiding with pedagogically advanced embedded training tools make it possible to create a training device that comes very close to producing optimized training. Such a training device using "symbiotic technologies" would help push simulator training into the next generation Human Factors (Boff, 2006). Imagine a flight simulator that can be completely managed by a student pilot. The simulator will teach the student according to his/her current level of skill and mental state. On a bad day the student will receive easier tasks, on a good day more challenging tasks. Each day the simulator will challenge the student at just the right level so he/she will learn new tasks in a pace that optimizes the training goal, and minimizes the time required to achieve that goal. The simulator will register and understand the students' learning curve and always provide an optimal learning environment.

REFERENCES

Alfredson, J. (2007). *Differences in Situational Awareness and How to Manage Them in Development of Complex Systems* (Dissertation). Linköping University: Linköping, Sweden.

Alfredson, J. & Nählinder, S. (2002). Pilot's eye as a source of information for a future adaptive aircraft system. *Proceedings of the 34th Annual Congress of the Nordic Ergonomics Society*. Vol. 1, pp. 21-26.

Alfredson, J., Nählinder, S. & Castor, M. (2004). Measuring eye movements in applied psychological research - five different techniques - five different approaches. *FOI-R--1406--SE. FOI Methodology Report*: FOI: Linköping, Sweden.

Alfredson, J., Angelborg-Thanderz, M., van Avermaete, J., Bohnen, H.G.M., Farkin, B., Ohlsson, K., Svensson, E. & Zon, G.D.R. (1997). *Dynamic Measures of Pilot Mental Workload (PMWL), Pilot Performance (PP), and Situational Awareness (SA)* (Tech. Rep. No. VINTHEC-WP3-TR01). Amsterdam: National Aerospace Laboratory NLR.

Angelborg-Thanderz, M. (1990). *Military Flight Training at a Reasonable Price and Risk*, (Dissertation). Economics Research Institute, Stockholm School of Economics: Stockholm, Sweden.

Bell, H.H. & Waag, W.L. (1998). Evaluating the effectiveness of flight simulators for training combat skills: A review. *The International Journal of Aviation Psychology.* Vol. 8(3), pp. 223-242.

Berntson, G.G. & Stowell, J.R. (1998). ECG artifacts and heart period variability: Don't miss a beat! *Psychophysiology*. Vol. 35(1), pp. 127-132.

Boff, K.R. (2006). Revolutions and shifting paradigms in human factors & ergonomics. *Applied Ergonomics*. Vol. 37(4), pp. 391-399.

Borgvall, J., Castor, M., Nählinder, S., Oskarsson, P-A. & Svensson, E. (2008). Transfer of training in military aviation. *FOI-R--2378--SE. FOI Methodology Report.* FOI: Linköping, Sweden.

Boucsein, W. & Backs, R.W. (2000). Engineering psychophysiology as a discipline: historical and theoretical aspects. In R.W. Backs & W. Boucsein (eds.), *Engineering Psychophysiology - Issues and Applications*. pp. 111-138. Lawrence Erlbaum Associates: Mahwah, NJ, USA.

Bürki-Cohen, J. (2003). Evidence for the need of realistic radio communications for airline pilot simulator training and evaluation. In *Proceedings of the International Conference Simulation of the Environment*, Royal Aeronautical Society, 5-6 November, London, UK.

Bürki-Cohen, J., Boothe, E.M., Soja N.N., DiSario, R., Go, T. & Longridge, T. (2000). Simulator fidelity - the effect of platform motion. In *Proceedings of the International Conference Flight Simulation – The Next Decade*, 10-12 May 2000. Royal Aeronautical Society: London, UK.

Castor, M., Hanson, E., Svensson, E., Nählinder, S., Le Blaye, P., MacLeod, I., Wright, N., Alfredson, J., Ågren, L., Berggren, P., Juppet, V., Hilburn, B. & Ohlsson, K. (2003). *GARTEUR Handbook of Mental Workload Measurement*. Group of Aeronautical Research and Technology in Europe Technical paper 145.

Dahlström, N., & Nählinder, S. (2006). A comparison of two recorders for obtaining in-flight heart rate data. *Applied Psychophysiology and Biofeedback*, Vol. 31(3), pp. 273-279.

Dahlström, N., & Nählinder, S. (2009). Mental Workload in Simulator and Aircraft during Basic Civil Aviation Training. *International Journal of Aviation Psychology*. (IN PRESS).

Dahlström, N., Nählinder, S., Wilson, G.F., & Svensson, E. (2009). Recording of Psychophysiological Data during Aerobatic Training. *International Journal of Aviation Psychology*. (ACCEPTED FOR PUBLICATION).

Eggemeier, F.T., Biers, D.W., Wickens, C.D., Andre, A.D., Vreuls, D., Billman, E.R. & Schueren, J. (1990). Performance assessment and workload evaluation systems: Analysis of candidate measures, *Technical Report No. HSD-TR-90-023*, Brooks Air Force Base, TX: Armstrong Aerospace Medical research Laboratory.

Endsley, M.R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors*, Vol. 37(1), pp. 32-64.

Fahrenberg, J. & Wientjes, C.J.E. (2000). Recording methods in Applied Environments. In R.W. Backs & W. Boucsein (eds.), *Engineering Psychophysiology - Issues and Applications*. pp. 111-138. Lawrence Erlbaum Associates: Mahwah, NJ, USA.

Gopher, D. & Donchin E. (1986). Workload – An examination of the concept, in Boff, K.R. & Kaufman, L. & Tomas, J. (eds.), *Handbook of Perception and Performance, Vol II: Cognitive processes and performance.* pp. 41-1 – 41-49. Wiley: New York.

Jöreskog, K.G. & Sörbom, D. (1984). *LISREL VI. Analysis of Linear Structural Relationships by Maximum Likelihood, Instrumental Variables, and Least Squares Methods.* Department of Statistics: University of Uppsala, Sweden.

Hancock, P.A. & Diaz, D.D. (2002). Ergonomics as a foundation for a science of purpose. *Theoretical Issues in Ergonomics Science*. Vol. 3(2), pp. 115-123.

Hankins, T.C. & Wilson, G.F. (1998). A comparison of heart rate, eye activity, EEG and subjective measures of pilot mental workload during flight. *Aviation Space and Environmental Medicine*, Vol. 69(4), pp. 360-367.

Hart, S.G. & Wickens, C.D. (1990). Workload assessment and prediction, in H.R. Booher (ed.), *MANPRINT: An Approach to Systems Integration*. van Nostrand Reinhold: New York, NY.

Hays, R.T., Jacobs, J.W., Prince, C. & Salas, E. (1992). Requirements for Future Research in Flight Simulation Training: Guidance Based on a Meta-Analytic Review. *The International Journal of Aviation Psychology*, Vol. 2(2), pp. 143-158.

Heiman, G.W. (2001). *Understanding Research Methods and Statistics*. Houghton Mifflin: Boston, USA.

Hewson, D.J., McNair, P.J. & Marshall, R.N. (1999). Aircraft control forces and EMG activity: Comparison of novice and experienced pilots during simulated take-off and landing. *Aviation Space and Environmental Medicine*, Vol. 70(8), pp. 745-751.

Hutchins, E. (1995). *Cognition in the Wild*. The MIT Press: Cambridge, MA.

Jorna, P.G.A.M. (1993). Heart rate and workload variations in actual and simulated flight. *Ergonomics*, Vol. 36(9), pp. 1043-1054.

Keppel, G. (1991). *Design and Analysis. A Researcher's Handbook.* Prentice-Hall, Inc: Upper Saddle River, NJ.

Kneebone, R. (2003). Simulation in surgical training: educational issues and practical training. *Medical Education*. Vol. 37(3), pp. 267-277.

Lee, A.T. (2005). *Flight Simulation Virtual Environments in Aviation*. Ashgate Publishing Limited: Aldershot, England.

Lee, Y-H. & Liu, B-S. (2003). Inflight workload. Assessment: comparison of subjective and physiological measurements. *Aviation, Space and Environmental Medicine.* Vol. 74(10), pp. 1078-1084.

Longridge, T., Bürki-Cohen, J., Go T.H. & Kendra, A.J. (2001). Simulator Fidelity considerations for training and evaluation of today's airline pilots. Paper presented at *the 11th International Symposium on Aviation Psychology*, Columbus, OH, USA.

Magnusson, S. & Berggren, P. (2002). Dynamic assessment of pilot mental status. In *Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting*. (Baltimore, MD). Human Factors and Ergonomics Society: Santa Monica, CA.

Magnusson, S. (2002). Similarities and differences in psychophysiological reactions between simulated and real air-to-ground missions. *International Journal of Aviation Psychology*. Vol. 12(1), pp. 49-61.

Magnusson, S., Berggren, P., Danielsson, B. & Svensson, E. (2001). Dynamisk värdering av operatörsfunktion för framtida systemutveckling. *FOI-R--0430--SE. FOI Methodology Report:* FOI: Linköping, Sweden.

Nählinder, S., Berggren, P. & Persson, B. (2005). Increasing training efficiency using embedded pedagogical tools in a combat flight simulator. In *Proceedings of the Human Factors and Ergonomics Society. 49th Annual Meeting.* (Orlando, FL). Human Factors and Ergonomics Society: Santa Monica, CA.

Nählinder, S. (2004). ACES - Air combat evaluation system. *FOI-R--1368--SE. FOI Methodology report:* FOI: Linköping, Sweden.

Nählinder, S. (2006). A Human-Factors perspective on simulator fidelity assessement. *FOI-R--2047--SE. FOI Scientific Report:* FOI, Linköping, Sweden.

Nählinder, S., Berggren, P. & Svensson, E. (2004). Reoccurring LISREL patterns describing mental workload, situation awareness and performance. *Proceedings of the Human Factors and Ergonomics Society 48th Annual Meeting – 2004*, HFES New Orleans, LA. Santa Monica, CA: Human Factors and Ergonomics Society.

Nyssen, A.S., Larbuisson, R., Janssens, M., Pendeville, P. & Mayné, A. (2002). A Comparison of the training value of two types of anesthesia simulators: computer screen-based and mannequin-based simulators. *Anesthesia & Analgesia.* Vol. 94, pp. 1560-1565.

O'Donnell, R.D. & Eggemeier, F.T. (1986). Workload assessment methodology. In Boff, K.R. & Kaufman, L. & Tomas, J.P. (eds.), *Handbook of Perception and Performance, Vol II: Cognitive Processes and Performance.* pp. 42-1 – 42-49. Wiley: New York.

Perey, P. (2008). Future simulation technologies. *Proceedings of the Royal Aeronautical Society Annual International Flight Crew Training Conference*. Sept 24-25, 2008: London, UK.

Rencrantz, C., Lindoff, J., Svensson, E., Norlander, A., & Berggren, P. (2006). Interoperabilitets- och metodstudie i en operativ stridsledningscentral. *FOI-R--2040--SE. Scientific Report*. FOI: Linköping, Sweden.

Roscoe, A.H. (1987). In-flight assessment of workload using pilot ratings and heart rate. In A.H. Roscoe (ed.), *The Practical Assessment of Pilot Workload*, AGARDograph No. 282.

Roscoe, A.H. (1992). Assessing pilot workload. Why measure heart rate, HRV and respiration? *Biological Psychology*, Vol. 34(2–3), pp. 259–287.

Roscoe, A.H. (1993). Heart rate as a psychophysiological measure for in-flight workload assessment. *Ergonomics*, Vol. 36(9), pp. 1055-1062.

Rouse, B. (1988). Adaptive aiding for human/computer control. *Human Factors*. Vol. 30(4), pp. 431-443.

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. *International Journal of Aviation Psychology*. Vol. 8(3), pp. 197-208.

Sterman, M. B., & Mann, C. A. (1995). Concepts and applications of EEG analysis in aviation performance evaluation. *Biological Psychology*, Vol. 40(1), pp. 115-130.

Svensson, E., Angelborg-Thanderz, M. & Wilson, G. F. (1999). *Models of Pilot Performance for Systems and Mission Evaluation Psychological and Psychophysiological Aspects*. (U.S. Air Force Rep. No. AFRL-HE-WP-TR-1999-0215). Wright-Patterson Air Force Base: Ohio, USA.

Svensson, E., Angelborg-Thanderz, M., Sjöberg, L. & Olsson, S. (1997). Information complexity - mental workload and performance in combat aircraft. *Ergonomics*, Vol. 40(3), pp. 362-380.

Svensson, E. & Wilson, G.F. (2002). Psychological and psychophysiological models of pilot performance for systems development and mission evaluation. *International Journal of Aviation Psychology,* Vol. 12(1), pp. 95-110.

Talleur, D. A., Taylor, H. L., Emanuel, T. W., Rantanen, E. & Bradshaw, G. L. (2003). Personal Computer Aviation Training Devices: Their Effectiveness for Maintaining Instrument Currency. *International Journal of Aviation Psychology*. Vol. 13(4), pp. 387-399.

Taylor, H.L., Talleur, D.A., Emanuel, T.W. & Rantanen, E.M. (2005). *Transfer of Training Effectiveness of a Flight Training Device (FTD).* Paper presented at the International Symposium on Aviation Psychology, Columbus, OH.

Vaden, E.A. & Hall, S. (2005). The effect of simulator platform motion on pilot training transfer: A meta-analysis *International Journal of Aviation Psychology.* Vol. 15(4), pp. 1050-8414.

Wickens, C.D. & Hollands, J.G. (2000). Assessing Mental Workload. In C. D. Wickens (ed.), *Engineering Psychology and Human Performance. 3rd ed*., pp. 459-471. Prentice-Hall: Upper Saddle River, NJ.

Wilson, G.F. & Russell, C.A. (2003). Operator functional state classification using multiple psychophysiological features in an air traffic control task. *Human Factors*. Vol. 45(3), pp. 381-389.

Wilson, G.F. & Russell, C.A. (2007). Performance enhancement in an uninhabited air vehicle task using psychophysiologically determined adaptive aiding. *Human Factors*. Vol. 49(6), pp. 1005-1018.

Wilson, G.F. (1993). Air-to-ground training missions: A psychophysiological workload analysis. *Ergonomics*, Vol. 36(9), pp. 1071-1087.

Wilson, G.F. (2001). Psychophysiological inflight monitoring. In J. Fahrenberg & M. Myrtek (eds.), *Progress in Ambulatory Assessment.* pp. 435-454. Hogrefe & Huber Publishers: Seattle, WA.

Wilson, G.F. (2002a). Psychophysiological test methods and procedures. In S. G. Charlton & T. G. O'Brien (eds.), *Handbook of Human Factors Testing and Evaluation, 2nd ed.* pp. 127-156. Lawrence Erlbaum Associates: Mahwah, NJ, USA.

Wilson, G.F. (2002b). A comparison of three cardiac ambulatory recorders using flight data. *International Journal of Aviation Psychology*. Vol. 12(1), pp. 111-119.