Workload of Airport Tower Controllers: Empirical Validation of a Macro-cognitive Model

Hardy Smieszek*. Fabian Joeres. Nele Russwinkel****

**Technische Universität Berlin, research group prometei, Berlin, 10587
Germany (Tel: +49 30-314-29635; e-mail: hardy.smieszek@mailbox.tu-berlin.de).
** Technische Universität Berlin, Department of Cognitive Modeling in dynamic
human-machine-systems, Berlin, 10587 Germany (e-mail:
fabian.joeres@tu-berlin.de; nele.russwinkel@tu-berlin.de)*

Abstract: The paper focuses on a validation study for a macro-cognitive model of mental workload of airport traffic controllers (ATCOs). This model constitutes a means for quantifying and analysing the distribution of ATCOs workload over time dependent on different traffic scenarios. Workload is modelled via the amount of chunks in working memory. In a one-factor experimental Simulator-Study workload ratings were gathered with different traffic load, using a modified RSME scale within and NASA-TLX after each scenario. The data are analysed and discussed concerning the successful experimental manipulation and the estimation of goodness of fit of the model and experimental data. Therefore the average workload ratings of the participant and the model within each scenario as well as the distribution of workload over time for each participant-model pair are compared. The developed model can serve as a tool for the design of adaptive automation and supporting systems and can help to better understand workload dynamics of airport traffic controllers.

1. INTRODUCTION

The responsibility of traffic controllers (ATCOs) in airport control towers is the safe and efficient handling of aircraft movements within their designated control zone. Without permission of an ATCO (so called clearances), no aircraft or vehicle is allowed to take any action at the airfield. Therefore ATCOs constantly have to consider a trade-off between efficiency and thoroughness (Hollnagel, 2009). This is a highly complex and dynamic task which needs a high level of attention and experience. As human information processing is limited (Kramar & Spinks, 1991; Wickens, 1984, 1991), it is essential to consider these limitations most notably for safety reasons. This is especially critical considering the growing amount of air traffic and therefore growing amounts of take-offs, landings and taxi procedures controllers have to handle. Moreover, attempts to assign two smaller airports to only one controller (so called remote control cf. e.g. Fürstenau et al., 2009), constitute an indirect increase of air traffic. To adjust to these changes, the abovementioned trade-off has to be shifted towards efficiency, which can lead to a higher probability of failure and therefore a higher risk of accidents (Hollnagel, 2009). It is therefore important to identify periods of high task- and workload, when cognitive overload of the ATCO is likely to occur, in order to counteract efficiently. One possibility is the implementation of adaptive automation. To reach this goal, a macro-cognitive modelling approach was chosen that incorporates the environment (airport), the interaction (incoming and outgoing information) and the controller’s cognitive processes into one model. Also macro-cognitive models aim to simulate complex cognitive processes in realistic environments (Smieszek & Rußwinkel, 2013).In this vein the MATriCS (Model of Airport Traffic Control System) model was developed as a holistic modelling approach. It takes into account all phases of information processing with a special focus set on mental workload of ATCOs in the control tower at airports. With this approach it is possible to simulate the distribution of workload of the ATCO over time and therefore to identify periods of high workload in respect of the traffic that has to be managed. In order to validate the model, a study was conducted in August 2013 at the TU Berlin. The study aimed at collecting data from participants within a simulated air traffic control task and to compare these workload data with the model data within the same traffic scenarios. The paper focuses on this validation study and the fit between model and human data.

After giving an overview about the underlying theoretical models the modelling approach and the structure of the model are briefly described. Finally the study the results will be presented and discussed.

2. THEORETICAL BACKGROUND

In order to adequately model workload of ATCOs it is necessary to look at the construct of workload first.

2.1 Mental Workload in Air Traffic Control

Most of the research on workload in the air traffic control sector has been conducted with en-route controllers and their tasks. These findings are therefore not fully applicable to airport controllers. It is often tried to infer the amount of workload from objectively observable variables like the amount of air traffic or communication events. Such approaches only consider the external task load factors (denoted as ATC complexity). However, Koros, Della Rocco, Panjwani, Ingurgio, and D'Arcy (2006) present a model in which not only external factors like air traffic characteristics influence workload. It is argued that workload depends on multiple variables. The resulting experience of workload is therefore equally influenced by individual factors like capacity limits, experience, or cognitive strategy, as well as external factors like traffic load. To find a valid way of modelling ATCO workload, several psychologically based concepts of workload should be considered.

2.2 Theoretical Concepts of Mental Workload

The term workload is often used without providing any definition. This may be the case because there is no consistent and comprehensive theory of mental workload in literature. Manzey (1998) identifies two dominating theoretical approaches of mental workload: activation-based and attention-based models. Within activation-based theories it is assumed that information processing requires energy. Mental workload is directly connected to this psychophysical effort and is therefore directly legible from psychophysiological activation indicators like heart rate (HR) or blood pressure (Manzey, 1998; Ribback, 2003).

The attention-based models result from research concerning dual task performance. It is assumed that human information processing capacity is limited. This capacity is not sufficient to perform two tasks at the same time without performance decrements (Kramar & Spinks, 1991; Wickens, 1984, 1991). The relative capacity demand of a task determines the amount of mental workload (Manzey, 1998). It is further assumed that there is not only one single resource reservoir which all cognitive processes equally demand. Rather multiple resources are assumed for different stages of information processing (perception and central processing as well as motor reaction), different modalities (auditory, visual) and processing codes (visuospatial, categorical-symbolic) (Wickens & McCarley, 2008; Wickens, 1984, 1991, 2002).

Hockey (1997) connects activation-based and attention-based approaches as he assumes multiple limited resources as energetic concepts where effort ensures energy supply for mental operations. Through such integrative theories limited information processing resources are interpreted as energetic concepts and are therefore accessible for physiological detection. In this work the concept of mental workload concurrent with integrative theories was used.

2.3 Cognitive Limitations and Working Memory

According to Wickens (1984, 1991, 2002) limitations of the information processing system primarily stem from working memory limitations. There are numerous approaches and theoretical models concerning structure, functions and limitations of working memory. As for workload, there are several findings which support multi-component theorys of working memory (Baddeley & Hitch, 1974; Baddeley, 2000, 2012; Cowan, 1999; Oberauer, Süß, Schulze, Wilhelm, & Wittmann, 2000). Although these theories make exact assumptions about the structure and functions of the single components of working memory, there are no consistent assumptions about its limitations.

Early approaches have a much easier concept of working memory but make more specific assumptions about its limitations (Klapp, Marshburn, & Lester, 1983; Oberauer et al., 2000; Sanders & McCormick, 1993). Within these approaches solely one sinlge capacity-limited resource reservoir is assumed. It is the well-known work of Miller (1956) who assumes a capacity limit of five to nine chunks which can be stored in working memory simultaniously. Here chunking is a subsumption of several pieces of information into one unit (a chunk). It is assumed that the ability of building chunks of unrelated items is influenced by knowledge, experience, and training (Baddeley, 1990; Ericsson & Chase, 1982). As ATCOs are highly experienced it can be assumed that they are highly able to build up and maintain numerous air traffic control specific chunks. This assumption is supported by studies conducted with ATCOs. For example, Sperandio (1969) conducted a study in which, after real control sessions, he asked ATCOs to recall all aircraft they had controlled in the last hour. On average, ATCOs were able to remember ten aircraft (Sperandio, 1969). Also, Bisseret (1971) conducted a study in which he tested three different groups of ATCOs with different levels of experience: highly experienced controllers, controllers who had just passed the lowest level qualification test, and trainees with three to six months less experience than the recently graduated (Bainbridge, 1975). He found that highly experienced ATCOs were able to remember ten aircraft on average as well. Because of these studies an average upper limit of working memory capacity of ATCOs of ten chunks can be assumed. Though individual differences in working memory capacity do exist, they are beyond the scope of this approach as initially an appropriate method for modelling workload has to be found.

2.4 Cognitive modelling of mental workload

For modelling purposes it is important to include both the abovementioned types of factors (external and individual) into account in order to generate a holistic picture of the phenomenon. Koros, Della Rocco, Panjwani, Ingurgio, & D'Arcy (2003) identify 29 external factors (complexity factors) for airport traffic controllers and ask ATCOs to rate their importance. In this survey, the amount of air traffic was rated as most important. Based on the abovementioned literature, working memory limitations can be identified as a second major factor. In order to implement both these factors, a holistic, macro-cognitive modelling approach is required (see Cacciabue & Hollnagel, 1995; Smieszek, Manske, Hasselberg, Russwinkel, & Möhlenbrink, 2013; Smieszek & Rußwinkel, 2013). This approach allows an adequate modelling of cognitive processes as well as external factors such as the amount of aircraft. The model is implemented by using the abovementioned theories of workload and cognitive limitations.

3. THE MATriCS MODEL

The MATriCS model's purpose is to simulate the integrated behaviour of an ATCO and the environment he works in. Therefore, the model follows a macro-cognitive modelling paradigm. As a tool that generally allows the creation of process models independent of the modelled domain, Coloured Petri Nets (CPN) were chosen (Jensen & Kristensen, 2009). CPNs allow for the implementation of cognitive processes and external processes within one modelling tool. They are a specific kind of Petri Nets which can be described as a graphical programming language. CPN-models can be structured hierarchically, i.e. nets can consist of several subnets that interact with each other and exchange information. With this in mind, even complex systems can be divided in several sub-models, each of which stays relatively simple and easily comprehensible. A detailed introduction to coloured petri nets in particular and petri nets in general can be found in Jensen and Kristensen (2009), Jensen (1997) and Reisig (2010).

3.1 Basic structure

The MATriCS model is structured according to this paradigm and is divided into three sub-models: an airport model, an interaction model, and a controller model.

The airport model describes the traffic processes that occur in the controller's environment. It consists of two components: an invariable process logic component and a variable airport structure component. Following the assumption that all airports share some fundamental processing elements such as certain locations (e.g. runways and stands) and certain actions (e.g. aircraft landing or taxiing from one location to another), these elements are represented in the process logic. On the other hand, the airport structure contains information about how these elements are arranged and connected and thereby describes an individual airport's layout. This concept allows for relatively simple implementation of new airports without modifying the CPN structure. The airport model is further explained in Manske, Smieszek, Hasselberg and Möhlenbrink (2013).

To enable the controller model to interact with the airport, the interaction model defines several channels of visual and auditory information exchange. Visual information about the current airport state can be acquired by the controller model via far view onto the airport as well as through a radar screen. Furthermore, the interaction model provides radio communication for aircraft requests and the controller model's clearances. In addition, information about all currently relevant aircraft is available on flight strips. Flight strips are small paper strips, containing information about one aircraft each. These kind of information acquisition and communication are based on Tavanti (2006).

The controller model's structure is based on Hacker's (1986) Action Regulation Theory. Hacker (1986) defines five phases and components of human action regulation: a) goal setting; b) collection of information and orientation; c) generation of plans; d) decision for an action alternative; e) execution of action. It is assumed that the goal in this context is set by the work environment. The controller model's primary goal is a safe and efficient coordination of aircraft movements.

According to the remaining four phases of action regulation, the controller model is divided into the four sub-modules *perceive*, *plan*, *decide*, and *act*. Whereas the plan and act components have primarily auxiliary functions within the model, perception and decision making processes are crucial for the controller model's information processing. A detailed description of the two modules is beyond the scope of this paper but they are further explained in Smieszek et al. (2013; perception) and Smieszek and Joeres (2013; decision making). All four components are linked to a working memory component that represents the controller model's cognitive capacity and thereby operationalizes the model's cognitive workload.

3.2 Working memory model and mental workload estimation

As argued before, a general working memory capacity of ten chunks is assumed. Chunks are created whenever the controller model acquires external information (such as the information, which aircraft is currently on the runway) or generates internal information (such as the decision, which clearance is to be given to a specific aircraft).

One chunk contains different kinds of information. When, for example, the model checks if a certain taxiway section is currently free, the *far view* or *ground radar* interaction channel is utilized. After the information is acquired, a chunk is generated in working memory that contains information about

- the requesting aircraft for which the check was conducted,
- the checked taxiway section,
- the state of that section (free or occupied), and
- (if occupied,) the aircraft currently in that section.

Which information is stored in a chunk depends on the way it was created and on its purpose in the overall decision making process.

Therefore, all of the abovementioned sub-modules (*perceive*, *plan*, *decide*, and *act*) are linked to the working memory module. Whenever a chunk is created or deleted in one of those modules, the number of currently maintained chunks is updated.

As argued before, the utilisation of working memory capacities is assumed to represent the current level of workload. Accordingly, the percentage of utilized working memory (current number of chunks divided by the limit of ten chunks) is used as a measure for the model's mental workload at any given time.

4. VALIDATION STUDY

The modeller's task is not only to build models but also to connect these models to the real phenomenon under study (Bub & Lugner, 1992; Möhlenbrink, 2011). Therefore, it is necessary to generate experimental data to which the model data can be compared in order to gather information about the goodness of fit of the model.

4.1 Background and aim of the study

In order to collect realistic workload data, a simulator-study was conducted. The experimentally recorded data was compared against the model's data. Therefore several workload measures were gathered during the experiment: one- and multidimensional subjective, physiological and performance measures. This paper focuses on the subjective measures. As there are nearly no reliability values for any workload measure, reliability is rather derived from a consistent coherence between measure and workload from prior studies. Therefore as multi-dimensional measure the NASA-TLX (Hart & Staveland, 1986) was used. As one-dimensional measure a modified version of the RSME-scale introduced by Eilers, Nachreiner and Hänecke (1986) was used. A validation-study of the modified scale showed very good correlations with established measures (NASA-TLX and the original RSME-scale) (Kosicki, 2011). To gather the workload data, the participants had to handle four traffic scenarios of 15 minutes each. It was intended to replicate the task of ATCOs as realistically as possible within the simulation environment. Simultaneously a high correspondence between the real task and the modelled task had to be ensured.

4.2 Experimental Design

The study has a one-factor repeated measure within subjects design. As *independent variable* the amount of traffic to be controlled was manipulated (low and high traffic). For each level a repeated measure was conducted. Therefore the participants had to work four traffic scenarios of 15 minutes each. To avoid effects of scenario-sequence and learning, the scenarios were permuted across all participants. This results in 4! = 24 possible arrangements of scenarios. As *dependent variables* different measures of mental workload were gathered. During each scenario 15 RSME values were recorded (one measure every minute). At the end of each scenario the NASA Task Load Index (NASA-TLX) was used as a multi-dimensional measure of the subjects' workload perception.

4.3 Traffic Scenarios

For the simulation four traffic scenarios were generated, lasting 15 minutes each: Two with high, two with low traffic demands. According to collected data from Frankfurt airport in the year 2009 (Huber, 2012) high traffic amount was defined with 20 movements (20 aircraft) per 15 minutes (1.3 movements/minute). Low traffic amount was defined with 10 movements per 15 minutes (0.6 movements/minute). Furthermore, the appropriateness of the traffic loads was trialled in pre-tests. The assignment of conditions in the present study to the respective scenarios is depicted in Table 1.

Table 1: Scenarios and traffic load conditions

High traffic amount	Low traffic amount
Scenario 1	Scenario 2
Scenario 3	Scenario 4

As ATCOs' workload can also be influenced by the airport configuration and complexity (cf. Koros et al., 2003), all four scenarios took place at the same simple-structured airport consisting of only one runway and one main taxiway.

4.4 Estimation of goodness of fit and hypotheses

Statistical tests were conducted to check if the experimental task load manipulation was successful. For this, two hypotheses concerning the two subjective measures were formulated. In both it is assumed that a higher amount of air traffic leads to higher perceptions of workload and therefore to higher ratings in the subjective measures.

H1: Subjective ratings obtained with RSME are higher for high traffic load compared to low traffic load (a-d). Ratings of Scenarios with the same traffic amount do not differ (e-f)

H2: Subjective ratings obtained with NASA-TLX are higher for high traffic load compared to low traffic load (a-d). Ratings of Scenarios with the same traffic amount do not differ (e-f).

A pairwise comparison of each scenario median was conducted in order to verify the following hypotheses which apply for both methods of measurement:

a) The subjective ratings in scenario 1 are higher than in scenario 2.

b) The subjective ratings in scenario 1 are higher than in scenario 4.

c) The subjective ratings in scenario 3 are higher than in scenario 2.

d) The subjective ratings in scenario 3 are higher than in scenario 4.

e) The subjective ratings in scenario 1 and 3 do not differ.

f) The subjective ratings in scenario 2 and 4 do not differ.

Beyond these statistical results, Schunn and Wallach (2005) describe a number of so-called goodness-of-fit-measures which serve as a means to estimate how good a model fits to experimental data. They describe three stages of goodness of fit estimation: 1. a visual comparison of the distribution of model- and experimental data; 2. the calculation of measures of how well relative trend magnitudes are captured; 3. the calculation of measures of deviation from exact location. For calculation of relative trend magnitudes they recommend the use of Pearson's r and r^2. As the dependent measures at hand are not an interval or ratio scale (but ordinal scale) those measures cannot be used here. In this case, Schunn and Wallach (2005) recommend the use of rank correlation coefficients Spearman's ρ and Kendall's τ. This applies especially in cases where the experimentally measured data (RSME resp. NASA-TLX) is only loosely related to the dependent measure of the model (e.g. chunks in working memory). Even though Schunn and Wallach (2005) do not see any reason to prefer Kendall's τ over Spearman's ρ, both were calculated as Kendall's τ is more insensitive against outliners and more conservative (Newton, 2002) while ρ is more often reported in other research and therefore facilitates better comparability. The calculation procedure of Spearman's ρ is similar to Pearson's r. However, not the actual values are taken but their ranks.

Spearman's ρ is calculated as follows:

$$\rho = 1 - \frac{6 \sum_i d_i^2}{n \times (n^2 - 1)}$$

In this, *d* is the difference between the ranks of model and data values and *n* is the amount of value pairs.

Kendalls's τ is calculated based on a comparison of concordant and not concordant pairs of ranks:

$$\tau = \frac{C - D}{\sqrt{(C + D + T_x) \times (C + D + T_y)}}$$

In this, *C* describes the pairs of values that are concordant and *D* describes the pairs of values that are not concordant T_x and T_y describe ties in the variables *x* and *y* (e.g. model and experimental data). Based on these trend-related goodness-of-fit measures, a third hypothesis can be formulated which concerns the expected results of comparisons between model data and experimental data:

H3: In comparing the distribution of workload over the four scenarios a highly positive correlation (> 0.6) between the average subjective rating and the model is expected.

H4: In comparing the distribution of workload over time of each individual participant and of the model a positive correlation between the RSME ratings and the model data is expected.

Measures of deviation from exact location take into account that a model can fit the trends of the data very well but completely miss the exact locations (values) of the experimental data. Therefore, Schunn and Wallach (2005) describe a number of measures from which RMSE (root mean squared error) is most commonly used. For the data set at hand there are several issues with respect to these measures. At first, the scales for measuring mental workload are somewhat arbitrary and not equidistant. In such cases of ordinal data, averages of quantitative deviations from exact location are not meaningful (Schunn & Wallach, 2005). Also, with subjective workload measures different ratings can be given by the participants even though the same situation is rated. A second problem is that the dependent measure is somewhat arbitrary with respect to the model because both, experimental and model-data, are not measured on the same dimension. Therefore it is not possible and also not meaningful to calculate such measures of deviation from exact data location.

4.3 Sample

As it is highly difficult and cost-intensive to recruit real ATCOs for such a study, a lay sample was used. Nevertheless, a profound understanding of the air traffic control task and simulated air traffic control environment was ensured by means of intensive instruction and training. 24 subjects participated in the experimental simulation study. Participants' age ranged between 22 and 39 years (m = 26.95, sd = 4.69). The sample size was determined according to Bortz & Döring (2006, p. 615) who recommend 23 participants by means of a univariate repeated measures ANOVA. In order to ensure full permutation of scenario sequence, the sample size was set to 24 participants.

4.4 Simulation environment

As no high-fidelity simulator was available, simulation-software for private users had to be used. Röbig, König, and Hofmann (2010) compare several such software products in order to build a "low-cost tower simulator."

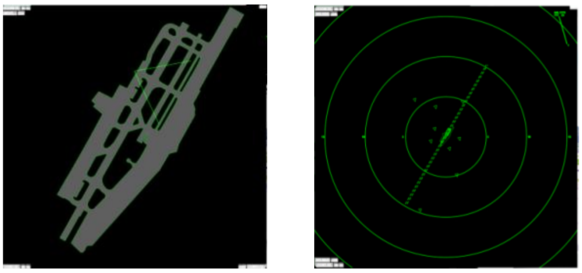

Figure 1: The two radar screens of the simulation environment the participants worked with (left ground radar; right air radar)

They describe the Simulation "Tower Simulator" (Wilco Publishing, 2008) as highly realistic. It provides the user with five different airports of different complexity. It further provides a ground and air radar, as well as a simulated far view, whereas the far views' graphic quality can be seen as outdated from today's point of view.

Requests of pilots are provided through simulated radio communication, communication from the user to the simulation works through input of text commands. As voice recognition could not be realised with the software, both requests from pilots and clearances by the participants were translated into the corresponding text commands for the simulation by the examiner.

In addition to the radar screen provided by the simulation (see figure 1), a strip bay (containing the relevant flight strips) as well as a headset for push-to-talk communication was provided to the participants. In order to not overcharge the participants and to warrant nearly realistic working conditions, an airport with low complexity was chosen.

4.5 Procedure

After participants' arrival and instructions about the general purpose of the study, demographic information was recorded in an according survey.

Detailed instructions on ATCOs' tasks and responsibilities in general, as well as on participants' specific tasks and the simulation environment followed. To ensure full understanding of the task, the participants were asked to summarize their responsibilities in their own words. Afterwards participants worked on a five minutes training scenario.

Pretests showed that several aspects of the complex task were particularly difficult for participants to understand and to execute. One such aspect was e.g. the estimation of the time an arriving aircraft needs until hitting the runway. Therefore participants often hesitated to issue a landing permission if the runway was still used by another aircraft. These aspects were repeated and emphasized after the training session.

Finally, the four abovementioned traffic scenarios were executed. At the end of each scenario (including the training scenario) the participants were given a five minute break. The whole experiment from participants' arrival to the end of the fourth traffic scenario lasted approximately 120 minutes.

5. RESULTS

As the data obtained from NASA-TLX and RSME is only ordinal data and not normally distributed the use of non-parametric tests is recommended (Bortz & Lienert, 2008; Fleid, 2009). Therefore, Friedmann's-ANOVA and post-hoc Wilcoxon signed rank tests with Bonferroni correction were used to calculate the statistical results. For estimation of goodness of fit, correlation coefficients Spearman's ρ and Kendall's τ were calculated. As explained earlier, estimations of deviations from exact data location is not informative for the measured construct at hand. That is why no such measures were calculated.

5.1 Experimental manipulation

It was assumed that for each measure of workload the ratings should be significantly higher in scenarios with high traffic amount compared to scenarios with low traffic amount. No difference was expected for scenarios with the same amounts of traffic. For the analysis of NASA-TLX only the subscale "mental effort" was analysed (NASA-TLX m. e.).

Friedmann-ANOVAs showed significant results for RSME and NASA-TLX m. e. median differences ($\chi2 = 33.734$, df 3, $p < 0.001$ for RSME; $\chi2 = 29.038$, df 3, $p < 0.001$ for NASA-TLX m. e.). Post-hoc Wilcoxon signed rank tests with Bonferroni correction ($\alpha = 0.0083$) showed that all alternative hypotheses can be confirmed. Table 2 shows all test results for NASA-TLX and RSME.

Table 2: Results for scenarios with different traffic loads

Scenario	RSME	NASA-TLX m. e.
1 vs. 2 (H1a, 2a)	T=25, p<0.001, r=0.496 **	T=24, p<0.001, r=-0.519 **
1 vs. 4 (H1b, 2b)	T=30, p<0.001, r=-0.495 **	T=6, p<0.001, r=-0.549 **
3 vs. 2 (H1c, 2c)	T=13, p<0.001, r=-0.565 **	T=37, p=0.001, r=-0.419 **
3 vs. 4 (H1d, 2d)	T=15, p<0.001, r=-0.557 **	T=44, p=0.001, r=-0.437 **

For testing the null hypotheses, the α-level was set to $\alpha = 0.20$ (Bortz, 2005) and no Bonferroni-correction was conducted to reduce the risk of a β-error. The post-hoc Wilcoxon signed rank test showed no significant difference for both scenarios in the NASA-TLX m. e. ratings and one scenario in the RSME rating. Solely for the comparison of the RSME rating of scenario 1 and 3 a small difference was discovered. The results are shown in Table 3.

Table 3: Results for scenarios with same traffic loads

Scenario	RSME	NASA-TLX m. e.
1 vs. 3 (H1e, 2e)	T=103.5, p(2-tailed) =0.190, r=-0.192 *	T=114.5, p(2-tailed)= 0.320, r=-0.146 **n.s.**
2 vs. 4 (H1f, 2f)	T=99, p(2-tailed)= 0.384, r=-0.192 **n. s.**	T=134.5, p(2-tailed)= 0.923, r=-0.015 **n.s.**

5.2 Average workload level

For estimating the average workload ratings for each scenario, medians (and average deviations from median MD) were calculated for each measure as they can just be seen as ordinal data. In contrast, the model data can be seen as scaled proportionally which is why the mean is reported here. The average workload ratings gathered with RSME and NASA-TLX m. e., as well as for the model are summarized in Table 4.

Table 4: Average workload ratings (median) and estimations of the model (mean) within each scenario

Scenario	RSME [%]	MD [%]	NASA-TLX m. e. [%]	MD [%]	Model [%]	sd [%]
1	25.06	10.30	63.54	20.24	33.92	6.31
2	16.52	10.15	34.51	19.38	12.28	3.57
3	29.39	11.08	50.93	21.40	24.56	5.20
4	17.29	9.85	31.47	20.41	8.14	2.38

With RSME a median workload rating of $M_1 = 25.06$ % (MD = 10.30 %) was reached for Scenario 1. With NASA-TLX the rating was much higher with $M_1 = 63.54$ % (MD = 20.24 %). The model showed a mean workload level of M = 33.92 % (sd = 6.31 %) was reached. The average workload ratings and estimations of the model are depicted graphically in Figure 2.

As it was noticed that the trend of RSME is different to NASA-TLX m. e., the total value of NASA-TLX as raw task load index (NASA-RTLX according to Byers, Bittner, & Hill, 1989) was calculated as well. This allows for detection of potential differences in diagnosticity of the single workload measures. It can be seen in Figure 2 that the absolute height of the workload ratings is slightly underestimated by the model but it approximately reaches the height reached by the RSME-ratings.

By calculating correlation coefficients over the four scenarios it can, nevertheless, be recognized, that the model reflects the

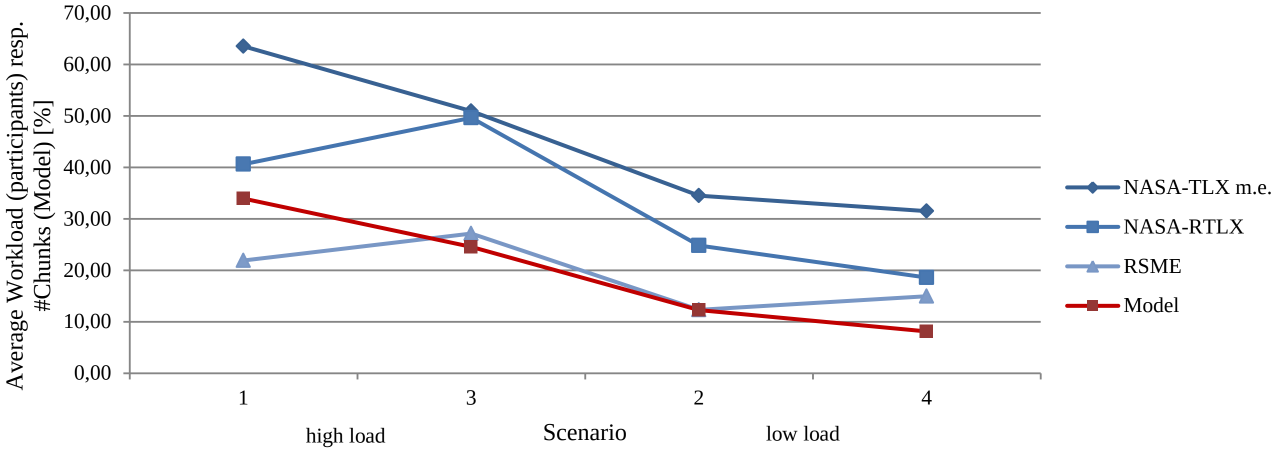

Figure 2: Average workload of the Model (#Chunks) compared to the average workload of RSME, NASA-TLX m.e. and NASA-RTLX [in %]

trend of the ratings achieved with the NASA-TLX m. e. scale with a correlation of ρ = 1,00. However the trend across the four scenarios of the NASA-RTLX and the RSME is replicated worse (ρ = 0,800 with NASA-RTLX; ρ = 0,600 with RSME). Instead the trends of the RSME and NASA-RTLX correlate better (ρ = 0,800). Both correlate worse with NASA-TLX m. e. (RSME: ρ = 0,600; NASA-RTLX: ρ = 0,800). These results are consistent with our hypothesis H3.

5.3 Distribution of workload over time

As a second measure of goodness of fit the workload distribution over time of the model was compared to the workload distribution over time of the participants. Therefore the participants had to rate their workload every minute during the scenarios via the RSME workload scale, such that for 15 points in time workload ratings could be gathered. These ratings were correlated with the workload estimation the model gave at exactly the same time.

Therefore, for each participant’s scenario a model simulation was conducted, such that data was generated for 4•24 = 96 scenarios (4 scenarios • 24 participants). For each of these 96 individual scenarios workload distribution over time was calculated for the model and the participants. Figure 3 depicts one such distribution for participant #24 in scenario one. It is important to note that the trend as well as peaks and valleys are replicated well by the model. It nevertheless seems as if the model more strongly reacts to differences in task load than the participant does, as the participant’s curve is flatter than the curve of the model. From the distributions in figure 3 a highly positive correlation can be expected as was assumed in hypothesis 4.

Because some participants’ ratings were missing, 64 out of 96 possible correlation coefficients could be calculated (66.67 %). Only five correlation coefficients were negative, which means that over 90 % were positive and compliant with our hypothesis. After Fishers-Z-transformation (Bortz & Schuster, 2010), an average correlation coefficient for each scenario as well as an overall average correlation coefficient was calculated. It is shown inTable 5.

It can be seen that an overall medium correlation could be reached with Spearman’s ρ (ρ = 0.427). The coefficient of Kendall’s τ is a little lower (τ = 0.336) which is consistent with our hypothesis H4.

Table 5: Correlation Coefficients

Scenario	Kendall’s τ	Spearman’s ρ
1	0.392	0.518
2	0.243	0.305
3	0.323	0.410
4	0.382	0.416
Overall	**0.336**	**0.427**

6. DISCUSSION

6.1 Experimental manipulation

Results show that the experimental manipulation was successful as with higher traffic load, higher workload ratings were measured. It can nevertheless be seen that even when the overall traffic amount within two scenarios was the same, workload was estimated differently. The reason could be that workload does not only depend on the overall amount of traffic within one scenario, but also on the distribution of aircraft (and therefore the distribution of workload) over time within the scenario. Because of different distributions of aircraft, different scenario difficulties arise which cause different estimations of workload. Nonetheless, this confounding of variables cannot be solved. By increasing the air traffic it is always the case that more complicated situations arise. Moreover, the exact scenario development depends severely on the individual participant’s actions and reactions. The sequence in which clearances are given affects the following movement patterns and thereby the occurrence of possible conflict situations. Better experimental control could be achieved by improving the controllability of the frequency and time of aircraft conflicts. Nonetheless, in reality the number of aircraft is a variable which can be influenced more easily than the number of critical situations this is why the conducted study provides a higher accordance to real situations. Looking at the model data, similar

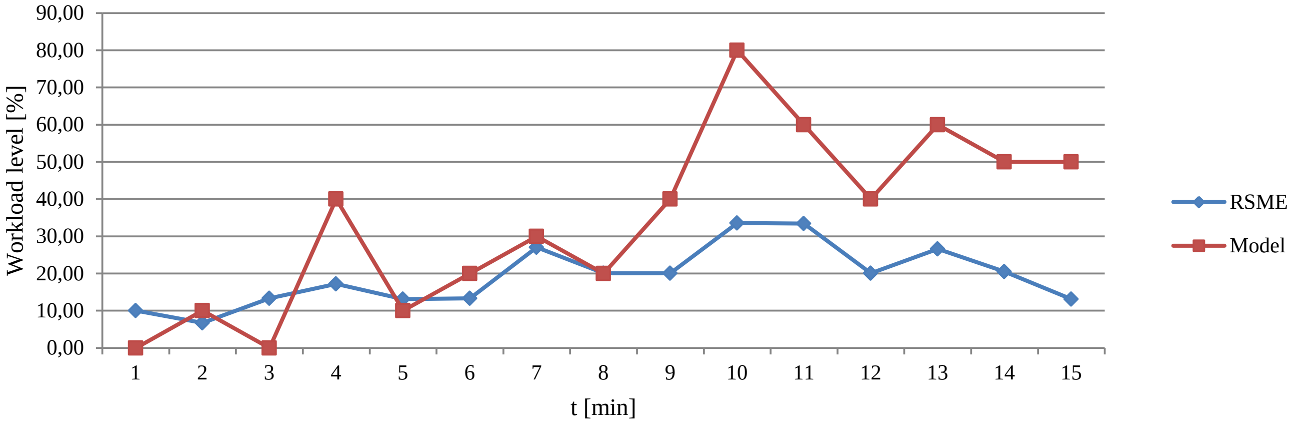

Figure 3: Distribution of Workload over time for participant 24 in Scenario 1

differences were detected. It can therefore be concluded, that the model not just detect differences in workload resulting from different traffic amount but also from different distributions of aircraft over time. Hence the model is able to estimate the amount of workload within a traffic scenario diagnostically.

6.2 Average workload level

It was shown that the mean workload level was usually underestimated by the model. Nonetheless, the model is similar to the by the RSME-ratings. The average workload ratings gathered with NASA-TLX m. e. are always much higher. It is possible that participants in hindsight rated a scenario as highly demanding using NASA-TLX while workload within the scenario (rated with RSME) was perceived rather moderate. This difference might also stem from the calculation of the average workload ratings. The calculated median of RSME consist of 15 measures taken during the scenarios (like with the model) while with NASA-TLX just one rating per scenario was gathered. It can also be assumed that with a rating in hindsight higher workload within a scenario is remembered more strongly than low workload.

It could further be shown that the average height of workload given by the model and recorded with NASA-TLX m. e. are highly correlated (ρ = 1.00). A lower correlation was reached between the model data and the data collected with the RSME. From the comparisons of RSME, NASA-TLX m. e. and NASA-RTLX it can be inferred that RSME is not as diagnostic for mental effort as initially assumed. Instead there are hints that with RSME, other factors affecting the subjective perception of workload (like time pressure) are somewhat included into participants' ratings. It might be the case that participants are not able to differentiate these factors within the demanding situation. Nevertheless, it could be shown that the model is highly diagnostic for mental effort as intended by the model development.

6.3 Distribution of workload over time

In examining the goodness-of-fit of the model using individual case assessments it can be seen that there is a positive relation between the model and the subjective ratings gathered with RSME-scale. There were only 5 negative correlation coefficients out of 64 (7.81 %). From the graphical depictions it can be assumed that, due to habituation and learning effects, ratings at the end of the scenarios are slightly lower which can explain negative correlations. The model is not able and not intended to take such effects into account. Furthermore, because of the pre-sorting of the flight strips, participants could have expected a high workload at the beginning of a scenario, which is then corrected during the scenario. Moreover, there were two participants who always marked one and the same anchor-point of the RSME-scale, such that no or just little variations arose. This might occur because the ratings were allocated a lower priority than the main task. Also the participants' motivation was possibly not high enough to give valid estimates of their current workload.

In comparing the distribution of RSME-ratings and model data visually, an increasing difference between both measures arises as the scenario proceeds. This might occur because the participants expect the end of the scenario and therefore adapt their ratings downwards. Additionally, when several aircraft and requests arose at (nearly) the same time, they may have been handled in different sequence by the model and by the participants. In phases of very high traffic loads, some aircraft were not processed at all by some participants. These differences and the resulting differences in the specific traffic situation might also account for low or negative correlations in some cases.

Finally also a lower diagnosticity for mental effort of the RSME-scale (as indicated by the results of section 5.2 and 6.2) could account for low correlations, because other factors affecting workload are somewhat included into participant's RSME ratings.

7 CONCLUSION

This study showed that the presented MATriCS model can predict the dependence of subjectively perceived workload on external task load which was manipulated by means of traffic load. Moreover, the model can predict situation-dependent workload within one scenario. Further research must be conducted to examine why the workload estimations of some participants differed severely from the model's predictions as indicated by negative correlations. One possible explanation is the application of individual strategies. Another possible explanation is that, as shown in figure 3, expectation of a stressful scenario as well as expectation of the imminent end of the scenario might have influenced the participants' perceived workload in the scenario's beginning and end. These negative correlations, however, occurred for only a minority of trials.

For further validation of the model, a high fidelity simulation with real air traffic controllers as participants should be used. This allows for generation of even more realistic data and therefore to proof the generalizability of the model to real air traffic control environments.

However, it should be noted that the MATriCS model shows a weak, but significant correlation with experimental workload data as well as a good estimation of the absolute height of workload. It therefore can be seen as a promising approach to estimating ATCOs' workload in different situations.

REFERENCES

Baddeley, A. D. (1990). *Human Memory: Theory and Practice*. Hove: Lawrence Erlbaum Associates.

Baddeley, A. D. (2000). The episodic buffer: a new component of working memory? *Trends in Cognitive Sciences*, **4**(11), 417–423.

Baddeley, A. D. (2012). Working memory: theories, models, and controversies. *Annual Review of Psychology*, **63**, 1–29.

Baddeley, A. D., & Hitch, G. (1974). Working Memory. In G. A. Bower (Ed.), *Recent advances in learning and motivation, Vol. 8*. New York: Academic Press.

Bainbridge, L. (1975). Working Memory in Air-Traffic Control. Retrieved October 30, 2012, from http://www.bainbrdg.demon.co.uk/Papers/WMemory.html

Bisseret, A. (1971). Analysis of mental processes involved in air traffic control. *Ergonomics*, **14**(5), 565–70.

Bortz, J. (2005). *Statistik* (6th ed.). Berlin: Springer.

Bortz, J., & Lienert, G. A. (2008). *Kurzgefasste Statistik für die Klinische Forschung*. Heidelberg: Springer.

Bortz, J., & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler* (7th ed.). Berlin, Heidelberg: Springer.

Brickenkamp, R., Schmidt-Atzert, L., & Liepmann, D. (2010). *d2-R: Test d2-Revision Aufmerksamkeits- und Konzentrationstest*. Göttingen: Hogrefe.

Bub, W., & Lugner, P. (1992). Systematik der Modellbildung. In Verein Deutscher Ingenieure (Ed.), *Modellbildung für Regelung und Simulation* (pp. 1 – 43). Düsseldorf: VDI-Verlag.

Byers, J. C., Bittner, A. C., & Hill, S. G. (1989). Traditional and raw task load index (TLX) correlations: are paired comparisons necessary? In A. Mital (Ed.), *Advances in industrial ergonomics and safety, I* (pp. 481 – 485). London: Taylor & Francis.

Cacciabue, P. C., & Hollnagel, E. (1995). Simulation of Cognition: Applications. In J.-M. Hoc, P. C. Cacciabue, & E. Hollnagel (Eds.), *Expertise and Technology: Cognition & Human-Computer Cooperation* (pp. 43–54). Hillsdale, New Jersey: Lawrence Erlbaum Associates Ltd.

Cowan, N. (1999). An Embedded-processes Model of Working Memory. In A. Miyake & P. Shah (Eds.), *Models of Working Memory* (pp. 62 – 101). Cambridge: Cambridge University Press.

Eilers, K., Nachreiner, F., & Hänecke, K. (1986). Entwicklung und Überprüfung einer Skala zur Erfassung subjektive erlebter Anstrengung. Zeitschrift Für Arbeitswissenschaft, **40**(4), 215–224.

Ericsson, K. A., & Chase, W. G. (1982). Exceptional Memory. *American Scientist, 70*, 607 – 615.

Fleid, A. (2009). *Discovering Statistics using SPSS* (3rd ed.). London: SAGE Publications Ltd.

Fürstenau, N., Schmidt, M., Rudolph, M., Möhlenbrink, C., Papenfuß, A., & Kaltenhäuser, S. (2009). Steps Towards the Virtual Tower: Remote Airport Traffic Control Center (RAiCe). In *ENRI International Workshop on ATM/CNS*. Tokyo, Japan.

Hacker, W. (1986). *Arbeitspsychologie*. Berlin: VEB Deutscher Verlag der Wissenschaften.

Hart, S. G., & Staveland, L. E. (1986). *Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research*. Moffett Field, CA: NASA-Ames Research Center.

Hockey, G. R. J. (1997). Compensatory control in the regulation of human performance under stress and high workload: A cognitive-energetical framework. *Biological Psychology*, **45**, 73 – 93.

Hollnagel, E. (2009). *The ETTO Principle: Efficiency-Thoroughness Trede-Off – Why things that go right sometimes go wrong*. Farnham: Ashgate.

Huber, S. (2012). *Optimierung des Funktionsumfanges von Airport Moving Maps durch Analyse von Runway Incursions (Dissertation)*. Technische Universität Berlin.

Jensen, K. (1997). *Coloured Petri Nets - Basic Concepts, Analysis Methods and Practical Use Volume 1: Basic Concepts* (2nd ed.). Berlin, Heidelberg: Springer.

Jensen, K., & Kristensen, L. M. (2009). *Coloured Petri Nets: Modelling and Validation of Concurrent Systems*. Berlin Heidelberg: Springer.

Klapp, S. T., Marshburn, E. A., & Lester, P. T. (1983). Short-term memory does not involve the "working memory" of information processing: The demise of a common assumption. *Journal of Experimental Psychology: General*, **112**, 240 – 264.

Koros, A., Della Rocco, P. S., Panjwani, G., Ingurgio, V., & D'Arcy, J.-F. (2003). *Complexity in Air Traffic Control Towers: A Flied Study. Part 1. Complexity Factors (DOT/FAA/CT-TN03/14)*. Atlantic City: Federal Aviataion Administration.

Koros, A., Della Rocco, P. S., Panjwani, G., Ingurgio, V., & D'Arcy, J.-F. (2006). *Complexity in Airport Traffic Control Towers: A Field Study. Part 2. Controller Strategies and Information Requirements (DOT/FAA/TC-06/22)*. Atlantic City: Federal Aviataion Administration.

Kosicki, D. (2011). *Der Einfluss von Testszenarien auf die Bewertung von Usability*. Thesis: Philipps-Universität Marburg.

Kramar, A. F., & Spinks, J. (1991). Capacity views of human information processing. In J. R. Jennings & M. Coles (Eds.), *Handbook of cognitive psychology: Central and* *autonomic nervous system approaches* (pp. 179 – 249). New York: Wiley.

Manske, P., Smieszek, H., Hasselberg, A., & Möhlenbrink, C. (2013). Entwicklung eines generischen Flughafen-Modells für die effizientere Makrokognitive Modellierung des Mensch-Maschine-Systems der Flughafenverkehrs-kontrolle mit farbigen Petri-Netzen. In E. Brandenburg, L. Doria, A. Gross, T. Günzler, & H. Smieszek (Eds.), *Proceedings of the 10th Berlin Workshop on Human-Machine Systems: Foundations and Applications of Human-Machine Interaction* (pp. 497–504). Berlin: Universitätsverlag der Technischen Universität Berlin.

Manzey, D. (1998). Psychophysiologie mentaler Beanspruchung. In F. Rösler (Ed.), *Ergebnisse und Anwendungen der Psychophysiologie (Enzyklopädie der Psychologie), C, Serie I, Bd. 5)* (pp. 799 – 864). Göttingen: Hogrefe.

Miller, G. A. (1956). The Magical Number Seven, Plus or Minus Two: Some Limits on our Capacity for. *Psychological Review*, **62**, 81–97.

Möhlenbrink, C. (2011). *Modellierung und Analyse von menschlichen Entscheidungsheuristiken mit farbigen Petrinetzen (Dissertation)*. Braunschweig: TU Braunschweig.

Newton, R. (2002). Parameters dehind "nonparametric" statistics: Kendall's tau, Somers' D and median differences. *The Stata Journal*, **2**(1), 45 – 64.

Oberauer, K., Süß, H.-M., Schulze, R., Wilhelm, O., & Wittmann, W. W. (2000). Working memory capacity - facets of a cognitive ability construct. *Personality and Individual Differences*, **29**, 1017–1045.

Reisig, W. (2010). *Petrinetze: Modellierungstechnik, Analysemethoden, Fallstudien*. Wiesbaden: Vieweg + Teubner.

Ribback, S. (2003). *Psychophysiologische Untersuchung mentaler Beanspruchung in simulierten Mensch-Maschine-Interaktionen*. Universität Potsdam.

Röbig, A., König, C., & Hofmann, T. (2010). Entwicklung eines Low-Cost-Towersimulators zur Evaluation arbeitswissenschaftlicher Fragestellungen. In *USEWARE 2010 - Grundlagen, Methoden, Technologien* (pp. 67–76). Baden-Baden: VDI-Berichte 2009.

Sanders, M. S., & McCormick, E. J. (1993). *Human factors in engineering and design*. New York: McGraw-Hill.

Schunn, C. D., & Wallach, D. (2005). Evaluating Goodness-of-Fit in Comparison of Models to Data. In *Psychologie der Kognition: Reden and Vorträge anlässlich der Emeritierung von Werner Tack* (pp. 115 – 154). Saarbrücken: University of Saarland Press.

Smieszek, H., & Joeres, F. (2013). Prospective decision making in a macro-cognitive model of airport traffic control system (MATriCS) based on coloured petri nets. In E. Brandenburg, L. Doria, A. Gross, T. Günzler, & H. Smieszek (Eds.), *Proceedings of the 10th Berlin Workshop on Human-Machine Systems: Foundations and Applications of Human-Machine Interaction* (pp. 505–512). Berlin: Universitätsverlag der Technischen Universität Berlin.

Smieszek, H., Manske, P., Hasselberg, A., Russwinkel, N., & Möhlenbrink, C. (2013). Cognitive Simulation of Limited Working Memory Capacity Applied to an Air Traffic Control Task. In R. West & T. Stewart (Eds.), *Proceedings of the 12th International Conference on Cognitive Modeling*.

Smieszek, H., & Rußwinkel, N. (2013). Micro-cognition and macro-cognition: trying to bridge the gap. In E. Brandenburg, L. Doria, A. Gross, T. Günzler, & H. Smieszek (Eds.), *Proceedings of the 10th Berlin Workshop on Human-Machine Systems: Foundations and Applications of Human-Machine Interaction* (pp. 335–341). Berlin: Universitätsverlag der Technischen Universität Berlin.

Sperandio, J.-C. (1969). Les variations du partage des taches entre un operater et son assistant, en fonction de la charge de travail du systeme. *Bulletin Du CERP*, **18**, 81 – 98.

Tavanti, M. (2006). *Control Tower Operations: Roles Description*. Brétigny-sur-Orge Cedex, France: EUROCONTROL Experimental Centre.

Wickens, C. D. (1984). Processing Ressources in Attention. In R. Parasuraman & D. R. Davies (Eds.), *Varieties of Attention*. Orlando: Academic Press.

Wickens, C. D. (1991). Processing Ressources and attention. In D. Damos (Ed.), *Multiple task performance* (pp. 3 – 34). London: Taylor & Francis.

Wickens, C. D. (2002). Multiple resources and performance prediction. *Theore*, **3**(2), 157–177.

Wickens, C. D., & McCarley, J. S. (2008). *Applied Attention Theory*. Boca Raton: Taylor & Francis.

Wilco Publishing. (2008). Tower Simulator. Retrieved July 12, 2013, from http://www.towersimulator.com/