Investigating Intrusiveness of Workload Adaptation

Felix Putze
Cognitive Systems Lab
Karlsruhe Institute of Technology
Karlsruhe, Germany
felix.putze@kit.edu

Tanja Schultz
Cognitive Systems Lab
Karlsruhe Institute of Technology
Karlsruhe, Germany
tanja.schultz@kit.edu

ABSTRACT

In this paper, we investigate how an automatic task assistant which can detect and react to a user's workload level is able to support the user in a complex, dynamic task. In a user study, we design a dispatcher scenario with low and high workload conditions and compare the effect of four support strategies with different levels of intrusiveness using objective and subjective metrics. We see that a more intrusive strategy results in higher efficiency and effectiveness, but is also less accepted by the participants. We also show that the benefit of supportive behavior depends on the user's workload level, i.e. adaptation to its changes are necessary. We describe and evaluate a Brain Computer Interface that is able to provide the necessary user state detection.

Keywords

Workload adaptive assistance; Brain Computer Interface; User Study; Intrusiveness

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation (e.g. HCI)**]: User Interfaces

1. INTRODUCTION

Since the beginning of the century, researchers in Human-Computer Interaction (HCI) systematically strive to design natural and intuitive user interfaces. One important aspect to achieve this goal is the design of adaptive interfaces which can react appropriately to the user's affective or cognitive state. Doing so requires two components: A recognizer which is able to detect the manifestation of a certain user state and an interaction manager with an adaptation strategy to react to the detected state. An important category of user state recognizers are passive Brain Computer Interfaces (BCIs) which use Electroencephalography (EEG) to determine the mental state of a user, for example the mental workload of a user [14]. A workload BCI offers potential benefits for HCI, especially in professional environments where specialists have to operate complex, dynamic and demanding user interfaces under time pressure. Adaptive interfaces in such scenarios have a huge potential of supporting a user when their mental workload rises, for example caused by the difficulty or number of tasks or due to external distractions. If not enough cognitive resources are available to satisfy all those demands, performance will drop and mistakes will occur, potentially with dire consequences. An adaptive interface is able to remedy those consequences by simplifying the interface or by simplifying the task by providing support in form of (partial) automation when high workload is detected. In this paper, we describe a study which compares different strategies designed to support a user in a complex, interactive task in high workload situations. A main contribution of this work is that we do not only look at the objective benefits of supporting the user but also at usability side effects of adaptive automation, namely the level of intrusiveness of a support strategy.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
ICMI'14, November 12–16, 2014, Istanbul, Turkey.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-2885-2/14/11 ...$15.00.

2. RELATED WORK

The general feasibility of workload BCIs has already been documented in the literature. For example, Lei et al. [15] used EEG to assess driver's workload. They used a simulated driving task and a secondary n-back memory task. They showed significant effects of task difficulty and multitasking in power spectrum, e.g. power attenuation in the α-band with increased working memory load. Murata et al. [16] recognized different workload levels in a continuous matching task where workload was controlled by levels of task difficulty. They used wavelets to estimate spectograms of each block for workload classification. Berka et al. [2] used simulated combat scenarios with five levels of difficulty assigned. EEG band power features were used to classify workload levels on epochs of one second length. Dijksterhuis et al. [5] recorded data from 34 participants in a driving simulator. They used stages of different driving demand levels to induce different levels of workload. Common Spatial Patterns (CSPs) were applied to generate person-specific spatial EEG filters for feature extraction. Wang et al. [17] controlled workload using different levels of difficulty of the Multi-Attribute Task battery. The authors concentrated on cross-person classification with a hierarchical Bayesian model trained on data from multiple participants. They achieved a recognition accuracy of 80% for person-independent classification on a corpus of eight participants.

Workload BCIs have been used in the literature for the development of workload adaptive user interfaces. For example, Wilson et al. [18] evaluated workload recognition for the Multi-Attribute Task Battery. In high workload condition, two of the subtasks were turned off and the authors showed that this adaptation improved the performance for the remaining tasks drastically. Kohlmorgen et al. [13] showed that a real-time detection of workload can be successfully used to manage distractions while driving in real driv-

ing situations by suppressing certain tasks during high workload. Chen et al. [3] employed a similar adaptation mechanism to reduce distraction by cell phone notifications in high workload situations. In [19], participants controlled a remotely piloted aircraft in scenarios with variable difficulty levels. The authors successfully evaluated an adaptation mechanism which reduced task speed and memory load when high workload is detected. Christensen et al. [4] used a similar scenario. Their presented system providds support in the form of target highlighting. They showed significant improvements compared to the non-adaptive baseline, but only after the user trained with the system for two days. Bailey et al. [1] compared the effect of adaptive automation versus adaptable automation for two different tasks. The authors saw that performance increases and subjectively perceived workload decreased for the adaptive system which does not require user-initiated intervention. They also observed that for the adaptable case, users hesitate to manually activate automation even in situations in which they would benefit from doing so.

The studies on workload adaptive systems cited above all concentrate on objective performance metrics to document the success of adaption. However, less attention has been payed to subjective criteria of usability, namely user satisfaction. This is surprising because meta-studies [7, 11] indicated that objective and subjective usability criteria are only weakly correlated and sometimes even contradictory. One example of a more thorough usability evaluation of an adaptive interface (but outside the context of workload adaptation) was given by Gajos et al. [8]. They explored different adaptation strategies to promote functionality which was frequently or recently used in the past. They investigated both the perceived benefits as well as the perceived costs (e.g. caused by suboptimal choices of the system or its unpredictability) and showed that even slightly different strategies are located at different spots in the benefit-cost space. They also show that there is no general relationship between perceived benefit and user satisfaction.

In the present paper, we try to close the gap on subjective evaluation of adaptive interfaces. For this purpose, we focus on multiple usability aspects of adaptive automation, as acceptance of a support strategy is a critical factor for its success. In [10], Heger et al. documented that adaptive switching between different behavior styles based on the workload estimate of a BCI improved both effectiveness and user satisfaction. In that study, the design of the supportive behavior was straight-forward as it transformed the presentation style of the system and did not interfere with the task directly. In cases where explicit support of the user is desired, the designer of an adaptive strategy has to chose between several possible behaviors. We believe that one of the major factors which determines the acceptance of supportive behavior is its level of intrusiveness; a highly intrusive support might be able to provide more substantial help, but risks be rejected by its users. In the present work, we conduct a user study to investigate this trade-off to determine how a workload-adaptive system can optimally support the user. To our best knowledge, this is the first evaluation of workload adaptive interfaces which addresses side effects of adaptation like intrusiveness.

3. METHODOLOGY

The goal of the presented study is to systematically investigate support strategies with different levels of intrusiveness. We do so for a scenario which realistically reflects the demands of a complex task: A constant stream of inputs which require the user's attention, planning and decision making under time pressure and multitasking. Those are requirements typical for a profession like emergency call agent, air traffic controller or dispatcher.

3.1 Task Design

More concretely, the participants of the study are told to work as a dispatcher in a factory, allocating workers with different skills to different upcoming requests. Requests appear randomly in a list. Each request has a title (e.g. 'mechanical defect in factory 2') and a skill requirement. Each request is also assigned a timeout, which is unknown to the user, after which it disappears and is marked as failed. Each worker has a primary skill and a secondary skill. The user is instructed to assign workers which match the required skill with their primary or at least secondary skill. Only if neither is possible, an unskilled worker is still better than letting the request go unaddressed. The quality of the assignment also depends on the time elapsed since the arrival of the request (shorter is better). When the worker is assigned, the request is removed from the list and the worker is disabled and unavailable for a period of time. As the number of workers is limited, this creates situations in which the user has to decide whether to wait for the return of a worker or to perform a sub-optimal assignment immediately. Each run of the task lasts for 3.5 minutes and has total of 30 requests. We call this primary task the *Dispatcher Task* (**DT**). It requires constant visual attention, quick decision making and planning, as assigning one worker may take away crucial options to handle other tasks satisfactorily. The frequent occurrence of new tasks maintains a constant workload level over the course of one run.

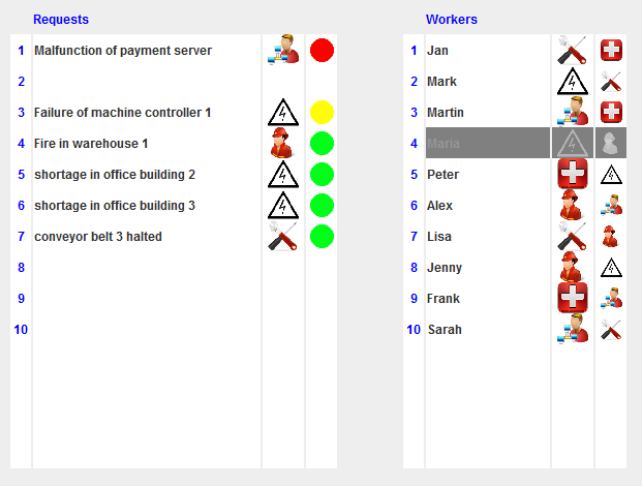

Figure 1: User Interface of the Dispatcher Task with list of requests on the left (with skill requirement and timer) and list of workers on the right (with primary and secondary skill). Workers marked in grey were already assigned and are not available.

In some configurations of the experiment, the participant is also handling an additional secondary task. This is a classification task which asks the user to sort out important messages from a constant flow of e-mails: The participant was asked to classify each message as either "spam" not "not spam" based on the occurrence of certain target words in subject line of the message. Entries which are not dealt with in a certain time window are discarded and scored as failures. This task runs in a second window on the same screen and requires the user to constantly divide their attention between the two tasks, making well-planned distribution of workers in the **DT** much more difficult. We call this secondary task the *Mail Task* (**MT**). We also refer to the configurations without **MT** as *low workload condition* and the configurations with **MT** as *high workload*

condition. We will justify this naming scheme later by comparing subjective workload ratings.

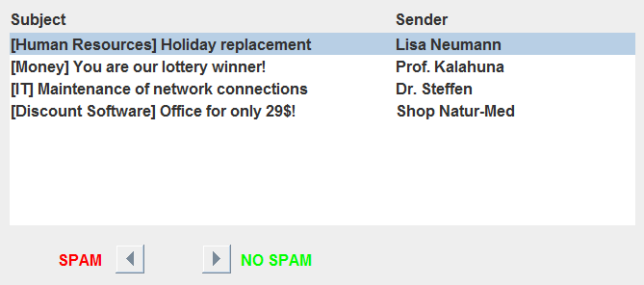

Figure 2: User Interface of the Mail Task, consisting of a list of e-mail headers which need to be classified as "spam" (red) or "no spam" (green).

3.2 Support Strategies

To relieve the participant in the high workload condition, we provide a virtual assistant which is able to intervene in the **DT** (e.g. by assigning workers to requests) and which can give advice and information to the user via synthesized speech. The behavior of the assistant is controlled by a *support strategy*.

We implement several support strategies. Those support strategies were designed to shift some of the decisions of the **DT** from the user to the system, thus reducing the number of required mental operations in a given time. The strategies differ mainly in their level of intrusiveness. More intrusive strategies are potentially more effective as they remove more decisions, but they also might reduce the level of control. Also note that the system is not omniscient concerning task timeouts and task requirements. More precisely, the system will perform optimal assignments only with a probability of 70%. If no such worker was selected (or none was available), the system selected a (random) worker with fitting secondary skill with a probability of 50%. If no such worker was selected (or none was available), it selected a (random) worker with no fitting skill. This procedure simulates the fact that in a typical use case in a professional environment, the user is an expert whose knowledge is difficult to reproduce automatically. It is instead approximated with an error prone heuristic. Therefore, the user will usually generate better results if no assistant is present and if they can fully concentrate on the task.

All support strategies are activated as soon as the number of present tasks passes a fixed threshold. The three (non-trivial) support strategies with decreasing level of intrusiveness are as follows: The `ACT` strategy, when activated, selects the oldest available task and assigns a worker to it. The strategy is constrained in that it may not act while the user has selected items to avoid interference with the user's decisions. The automatic assignment is then reported to user via speech synthesis with a statement that names the quality of skill match and the name of the removed task. This feedback is potentially useful as it informs them of the intervention and the resulting change in the task state. It also helps the user to judge whether dedicating more attention to the Dispatcher Task may improve assignment quality. However, ignoring the system statements does not influence the course of the task. The `OPT-OUT` strategy allows the user to exert more influence on the intervention of the system. Instead of directly executing an assignment, the system proposes it to the user by verbalizing it and by simultaneously marking it in a special color in the graphical interface. Again, those proposals are guaranteed to not interfere with a partial selection of the user. While the proposal is pending, the user can operate the task in the usual way. After five seconds, the proposal is executed (if still valid). The user has the ability to suppress this execution by pressing a button. In that case, the proposal is discarded and potentially replaced by a new one. The `OPT-IN` strategy further reduces the intrusiveness of the system. It generates and presents assignment proposals to the user in the same fashion as `OPT-OUT` does. The central difference is that for `OPT-IN`, the system requires a key press to accept and execute such a proposal. If not accepted within a certain period of time, the proposal is discarded and potentially replaced by a new one. For comparison, we call the 'support strategy' which does nothing and remains silent throughout a session `NONE`.

We made the design choice to use synthesized speech as output modality for all support strategies, as the visual load of both **DT** and **MT** is already high and because the participant can process verbalized information regardless of his focus of visual attention. While the strategies were mainly designed to support the user in a high workload situation, the system is not directly aware of the presence of a secondary task. If adaptive behavior is desired, we need to provide an external workload recognition.

3.3 Workload BCI

To provide adaptive behavior, the system has to recognize the user's workload level. The setup of the workload BCI mainly follows [10]: Classification is performed on windows of EEG data of 2s duration with an overlap of 1.5s. For preprocessing, the influence of ocular artifacts is reduced by performing Independent Component Analysis and automatically removing components containing artifact patterns. Frequency-based features (power of 28 frequency bins from 4 Hz to 45 Hz) are calculated for each electrode and then concatenated into a feature vector of 896 dimensions. A classifier based on Linear Discriminant Analysis is trained person-dependently on the binary classification task of separating low and high workload conditions. Temporal integration of the results of ten subsequent windows is performed to improve recognition stability. In contrast to [10], we perform an offline workload classification in this study, as we focus on the evaluation of the different support strategies.

3.4 Experimental Setup

The experimental setup is as follows: First, the participants were introduced to the Dispatcher Task with written and oral instructions. Following that, they performed several training runs of the task to familiarize with the keyboard layout and the task flow. Then, the order of the four support strategies was determined randomly and a block of four runs in that order was executed. Immediately before a run, the corresponding strategy was explained and demonstrated. After each run, the participant filled out a questionnaire on the user experience in that session (see Table 1). The questionnaire covered several aspects of user satisfaction which we deemed relevant for assessing the quality of a task. It included statements on subjective task performance, quality aspects of the system behavior, attribution of success and intrusiveness. The questionnaire used a five point Likert scale. Subjective workload was estimated using the NASA TLX questionnaire [9]. After the first block, the Mail Task was introduced and demonstrated. The participants performed a training run of the high workload condition. Again, the order of the four support strategies was determined and the participants performed a second block of four runs with subsequent questionnaires. In total, a participant performed eight runs plus two training runs, each with a duration of 3.5 minutes. This lead to a total recording time of 35 minutes for each participant. With

this setup, we recorded a total of 16 participants. Participants were all university students or staff members. Mean age of participants was 22.1 ($\sigma = 3.6$); Five of the participants were female. Participants were paid 15 Euro for their participation in the study. 12 of those sessions were performed with EEG recordings for the analysis of a BCI for the classification of mental workload. EEG was recorded using a 32 channel BrainVision actiCap with active electrodes, sampled at 500 Hz and referenced at Pz.

Item	Text
Q1	I had no problems with handling the task.
Q2	I am content with my performance.
Q3	I pressed keys randomly.
Q4	I was in control of the task.
Q5	The assistant supported me.
Q6	I could work relaxedly.
Q7	I listened carefully to the assistant.
Q8	The assistant helped in a timely fashion.
Q9	I felt patronized by the assistant.
Q10	The assistant distracted me from the task.
Q11	I felt I was not up to the task.
Q12	The assistant allowed accurate task execution.
Q13	The assistant behaved intrusive.
Q14	The assistant allowed fast task execution.
Q15	I wanted to succeed without support.
Q16	Task success was on me.
Q17	It was pleasant to work with the assistant.
Q18	I had to work against the assistant.
Q19	I would chose to work with the assistant.

Table 1: Items of the user satisfaction questionnaire for the DT.

4. EVALUATION

For evaluation, we looked at three research questions: First, do the support strategies lead to an improved task performance and is there a difference in the extend how they do so? Second, how are different support strategies assessed subjectively by the participants? Third, do benefit and subjective rating of support strategies change with the workload condition? If this is the case, we can make a strong point for the application of adaptive system behavior, i.e. switching between support strategies, depending on the detected workload level.

4.1 Objective Performance Metrics

Before analysis, we filter the data corpus by sorting our participants which are *high performers*, who are able to handle the Dispatcher Task with a success rate of 100% even in the presence of the Mail Task and their Mail Task success rate is also $> 95\%$. For those participants, no assistance of any type can improve their performance and this will also influence their subjective assessment of the support strategies. Of the 16 participants, five fit the definition of a high performer. For the analysis of task performance and user satisfaction, we exclude those participants to avoid ceiling effects. From exploratory sessions, we estimated that high performers exhibit the same effects as outlined below for a higher task difficulty.

We employ three different performance metrics for the **DT** and one performance metric for the **MT**: For both, we measure success rate (*SR*) as the relative number of items that was handled before they expired. For the **DT**, we additionally evaluate reaction time (*RT*) as the time it took to deal with an item (only regarding items which were eventually dealt with at all) and assignment quality (*AQ*) as the average match between assigned worker skill and request requirements (assignment quality of 2 means a primary skill match, assignment quality of 1 is a secondary skill match and assignment quality of 0 is no match).

		DT *SR* [%]	**DT** *RT* [s]	**DT** *AQ* [Quality]	**MT** *SR* [%]
DT only	NONE	0.91	7.18	1.64	-
	OPT-IN	0.86	7.65	1.52	-
	OPT-OUT	0.93	7.04	1.63	-
	ACT	0.99	5.15	1.67	-
DT+MT	NONE	0.79	8.95	1.39	0.93
	OPT-IN	0.88	8.35	1.47	0.92
	OPT-OUT	0.87	8.30	1.48	0.95
	ACT	0.99	5.61	1.67	0.96

Table 2: Average performance metrics for Dispatcher and Mail Task.

Table 2 summarizes averaged performance metrics for all eight runs. We first note that, unsurprisingly, the mail task has a strong impact on the performance in the Dispatcher Task: The average success rate drops significantly[1] by 13.2% ($t = 3.23$, $p = 0.004$) between the low workload condition and the high workload condition for theNONE strategy; the average skill match drops significantly by 15.2% ($t = 3.13$, $p = 0.005$) and the average reaction time rises significantly by 24.7% ($t = 2.38$, $p = 0.02$). Furthermore, the raw subjective workload index (unweighted sum of all TLX items) rises significantly by more than 30% ($t = 3.97$, $p = 0.001$).

In the low workload condition, performance is not influenced much by the employed strategies. There is a maximum improvement in *SR* of 8.2% for ACT compared to NONE, but also a decrease of 5.5% for OPT-IN compared to NONE. We see the same pattern for *AQ*. For *RT*, only ACT provides a substantial reduction compared to NONE. This is understandable as ACT is the only strategy where multiple assignments can be processed truly in parallel. Overall, the benefit of all support strategies is small. In the high workload condition, the gain of employing a support strategy is more substantial: *SR* now improves by 24.9% for ACT compared to NONE and overall, all strategies yield significantly higher *SR* than NONE ($t = 3.13$, $p = 0.005$ for ACT, $t = 3.24$, $p = 0.004$ for OPT-IN and $t = 2.11$, $p = 0.03$ for OPT-OUT). While in the low workload condition, there is no notable difference in *AQ*, in the high workload condition, there is an improvement in *AQ* of more than 20% for ACT compared to NONE ($t = 3.06$, $p = 0.006$). This means that under high workload, it is not only possible to handle more items, but the performed assignments are also better. This is an interesting result as the assistant is programmed to perform sub-optimally compared to an expert user. Also for OPT-IN and OPT-OUT, we see small improvements in *AQ* between 5% and 6%. *RT* is again only influenced positively by the ACT strategy. Overall, we see substantial improvements for the ACT strategy and positive effects for all three strategies. ACT increases all performance metrics to or above the levels of the low workload condition with the NONE strategy. Differences in task performance are also reflected in subjective workload assessment, measured by the TLX questionnaire. For example, the important item 'mental demand' drops significantly from 15.8 to 13.8 between NONE and ACT in the high workload condition ($t = 2.11$, $p = 0.03$) and there is

[1]For all differences reported as 'significant' in this section, this refers to a paired, one-sided t-test with $\alpha = 0.05$. Tests were family-wise error-corrected for multiple testing using the Bonferroni-Holm method. We report the t-value and the resulting p-value for each test.

a similar trend for the other TLX items. On the other hand, the OPT−IN strategy basically shows no difference in those dimensions compared to ACT. For the low workload condition, OPT−IN and OPT−OUT even increase the subjective workload index as they impose additional decisions on the user.

Id	Loading Items (Correlation)	Interpretation
F1	Q8 (0.87), Q12 (0.78), Q14 (0.76)	objective benefit of assistant
F2	Q9 (0.88), Q13 (0.93), Q18 (0.61)	intrusiveness of assistant
F3	Q2 (0.96), Q4 (0.65), Q6 (0.50)	task control
F4	Q10 (0.73), Q15 (0.69), Q16 (0.51), Q18 (0.52)	desired independence
F5	Q1 (0.92), Q3 (-0.60), Q7 (0.56)	overload

Table 3: Result of the factor analysis for user satisfaction questionnaire items.

4.2 User Satisfaction Metrics

Next, we analyze the answers for the questionnaires to evaluate how participants judge the different supporting strategies. A general summary of this judgment is given by the overall acceptance (Q19) of the strategies. Compared between runs with and without **MT**, acceptance for all three strategies improves for the high workload condition. This is explained by participants reporting that they are less ambitious to handle the task completely on their own in the high workload condition compared to the low workload condition (agreement to Q15 drops by 30.1% averaged across all strategies). This result is highly relevant for the application of adaptive user interfaces as it shows that a supportive behavior must not simply be activated all the time to be helpful, but must adapt to the user's workload level. The ranking of the strategies derived from acceptance: OPT−IN is significantly preferred to OPT−OUT ($t = -2.72$, $p = 0.006$) which is preferred to ACT (although the difference is less pronounced and therefore not significant: $t = -1.02$, $p = 0.16$). This means the order of preference is reversed compared to the order derived from performance improvement. Such discrepancy between objective performance metrics and subjective user satisfaction is long known in usability research [7]. To our knowledge, our study documents such relationship between objective and subjective usability metrics for the first time in the context of adaptive automation. When taking a more detailed look at the items of the questionnaire, we see that those cover different aspects of the interaction. To group those items, we perform an explorative factor analysis with Varimax rotation on all items except Q19. We extracted five factors, which in total explain more than 70% of the variance in the data. The resulting factors are summarized in Table 3. We see that the items of the questionnaire are grouped together in an interpretable way, representing objective benefit of the assistant, its intrusiveness, the amount of control exerted by the participant, the desired level of independence and the level of experienced overload. This result shows that there are actually different independent aspects of user satisfaction covered by the questionnaire. To discern the effect of workload conditions on the different factors and to analyze the influence of the factors on overall acceptance, we need to investigate them separately.

Table 4 presents the item scores for the different strategies, averaged across the items loading on the corresponding factors. We see significant differences between strategies for F1, F2 and F3. In contrast, differences for F4 and F5 are not significant. This means that participants perceive the objective benefits (F1) of ACT compared to OPT−OUT ($t = -1.76$, $p = 0.047$) and of OPT−OUT compared to OPT−IN ($p = 0.09$). On the other hand, they also evaluate ACT as much more intrusive (F2) compared to OPT−OUT ($t = 4.47$, $p = 0.0007$) which is perceived as slightly more intrusive than OPT−IN (however not significantly: $t = 0.68$, $p = 0.25$). This indicates that participants perceive task execution by the system itself as more intrusive, not the mode of presentation. Furthermore, participants also felt more strongly that they lost control over the task (F3) from OPT−IN to OPT−OUT ($t = 3.66$, $p = 0.001$) and from OPT−OUT to ACT ($t = 3.31$, $p = 0.002$). F4, i.e. the desire for independence from the assistant, on the other hand does not vary with strategy and seems to be a more stable personality trait. Experienced overload (F5) also does not change with the strategy. This is in line with the observation that none of the six workload dimensions of the NASA TLX correlates significantly with acceptance ($r \leq 0.19$ for all dimensions).

As the overall acceptance item Q19 was excluded from factor analysis, we can now predict this item from the resulting factors. For this purpose, we estimate a linear regression model with Q19 as dependent variable and F1, ..., F5 as independent variables. In the result, the overall model achieves an r^2 of 0.51. Looking at individual factors, F2 and F4 are significant predictors of Q19 ($p = 0.0001$ and $p = 0.04$, respectively) and most strongly influence the overall acceptance of the system, while the influence of the objective benefits of the assistant (i.e. F1) is not significant. Of those influential factors, F2 varies strongly between OPT−IN and the other two strategies, and is slightly more positive for OPT−OUT compared to ACT, but not significantly. This explains the observed preference pattern reflected by Q19, which behaves analogously. The fact that F4 ("'desired independence") is also a predictor of acceptance indicates that not only situational workload plays a role for strategy acceptance, but also the user's personality.

	F1	F2	F3	F4	F5
OPT−IN	2.80	3.53	3.92	2.83	2.78
OPT−OUT	3.01	2.94	3.17	2.83	2.52
ACT	3.41	2.82	2.41	2.92	2.88

Table 4: Scores for the different support strategies summarized for the factors from Table 3

4.3 Workload BCI Evaluation

We finally evaluate the recognition accuracy of the workload recognizer. We assign to all data from the low workload conditions of one participant the label LOW and assign to all data from the high workload condition the label HIGH. For this binary classification problem, we perform the analysis offline in a person-dependent 16-fold cross-validation. Overall, we see a recognition accuracy of 75.8%, with an standard deviation of 16.8%. If we exclude one participant for which technical problems compromised data quality of some electrodes, recognition accuracy improves to 79.3% with a standard deviation of 12.1%. This indicates that the system is able to reliably differentiate between the different workload levels, with similar accuracy as reported in related work (see first Section). Given that the training material is very heterogeneous, this is a satisfying result. [10] showed that a recognition accuracy in a similar range already allows adaptive automation with substantial improvements in task performance and user satisfaction compared to non-adaptive systems.

In a seminal work, Fairclough [6] dealt with fundamental issues of physiological computing. Following this work, it is important to reflect about the "specificity of the psychophysiological inference" and the "psychophysiological validity" of the workload recognizer. Concerning specificity, the investigated workload recognizer is limited by the fact that it is trained on only one combination of primary and secondary task. However, when we studied related works

(see first Section), we saw that frequency-based EEG features react specifically to high workload levels for a variety of different task selections. We can therefore conclude that despite a limit evaluation data set, the learned ability for class discrimination is specific to workload. One limitation of the present workload evaluation regarding psychophysiological validity is the fact that the order of the workload conditions is fixed (in contrast to the order of the support strategies). This can potentially bias the workload classifier towards learning temporal effects instead of workload differences. However, Jarvis et al. [12] showed that the employed workload classifier was robust against such ordering effects. As the classification setup in [12] was similar to the present one, we are optimistic that this robustness also transfers to the workload classification in the present work. Note that this ordering limitation will not substantially influence the relative behavioral results of the study. In general, participants improve their skills on both **DT** and **MT**, i.e. switching the order of workload conditions would only emphasize the benefit of the support strategies which we showed in the evaluation.

5. DISCUSSION

There are three main points we conclude from this user study: First, it is technically possible to estimate the workload level of a user in the employed scenario. Second, supporting behavior is helpful to the user and generally accepted, but only in the high workload condition. This shows the importance of adaptive user interfaces which turn on assisting behavior only when required. Third, and most importantly, the level of intrusiveness is a major determinant of how well a specific support strategy is perceived. The objectively most successful support strategy was ranked low compared to a more acceptable, less intrusive alternative. The results give no easy indication on what strategy is optimal in the given scenario. A designer will have to decide whether an additional performance gain is worth the cost of discontented users. This decision will for example be driven by the costs of mistakes during the main task. A reliable workload classifier can help to only activate such intrusive support when necessary to reduce the negative effect on user satisfaction to a minimum. Although the employed main task was presented with a cover story of a dispatching scenario, it is abstract in nature so we expect results can be transferred to any task with the same main properties, i.e. 1) heterogeneous but frequent inflow of requests, 2) distraction by a secondary task and 3) the property that work can be freely and independently distributed between human operator and machine. Examples of such tasks are air traffic control or crisis management.

Besides the rather small sample size (as five participants had been excluded as high performer) which limits the power of statistical analysis, one limitation of this study is the fact that workload recognition is performed offline. We argue that previous research indicates that online workload recognition is feasible at performance levels which are sufficient to provide significant usability improvements. Still, an analysis with an online workload recognizer for selection of supporting behavior would introduce realistic errors. Another limitation is that the strategies we investigated in this work are limited to local decisions, i.e. each request is treated independently. This is helpful as it maximizes the flexibility of the system and allows immediate switching between different behaviors. As the inflow of future tasks cannot be predicted by the system or the user, this flexibility is needed. Still, in slightly different scenarios, long-term variants of the presented strategies (which would store the user's decision on desired support over a certain period of time) could provide a more efficient and less intrusive behavior.

6. REFERENCES

[1] Bailey, N. R., Scerbo, M. W., Freeman, F. G., Mikulka, P. J., and Scott, L. A. Comparison of a brain-based adaptive system and a manual adaptable system for invoking automation. *Human Factors: The Journal of the Human Factors and Ergonomics Society 48*, 4 (2006), 693–709.

[2] Berka, C., Levendowski, D. J., Ramsey, C. K., Davis, G., Lumicao, M. N., Stanney, K., Reeves, L., Regli, S. H., Tremoulet, P. D., and Stibler, K. Evaluation of an EEG workload model in an aegis simulation environment. In *Defense and Security*, vol. 5797 (2005), 90–99.

[3] Chen, D., and Vertegaal, R. Using mental load for managing interruptions in physiologically attentive user interfaces. In *Extended Abstracts on Human Factors in Computing Systems* (USA, 2004), 1513–1516.

[4] Christensen, J. C., and Estepp, J. R. Coadaptive aiding and automation enhance operator performance. *Human Factors: The Journal of the Human Factors and Ergonomics Society* (2013), 0018720813476883.

[5] Dijksterhuis, C., Waard, D. d., and Mulder, B. L. J. M. Classifying visuomotor workload in a driving simulator using subject specific spatial brain patterns. *Frontiers in Neuroprosthetics 7* (2013).

[6] Fairclough, S. H. Fundamentals of physiological computing. *Interacting with Computers 21*, 1–2 (2009), 133–145.

[7] Frøkjær, E., Hertzum, M., and Hornbæk, K. Measuring usability: are effectiveness, efficiency, and satisfaction really correlated? In *Proceedings of the Conference on Human Factors in Computing Systems* (New York, USA, 2000).

[8] Gajos, K. Z., Czerwinski, M., Tan, D. S., and Weld, D. S. Exploring the design space for adaptive graphical user interfaces. In *Proceedings of the Working Conference on Advanced Visual Interfaces* (New York, USA, 2006).

[9] Hart, S. G., and Staveland, L. E. Development of NASA-TLX (task load index): Results of empirical and theoretical research. In *Advances in Psychology*, vol. 52 of *Human Mental Workload*. North-Holland, 1988, 139–183.

[10] Heger, D., Putze, F., and Schultz, T. An EEG adaptive information system for an empathic robot. *International Journal of Social Robotics 3*, 4 (2011), 415–425.

[11] Hornb\a ek, K., and Law, E. L.-C. Meta-analysis of correlations among usability measures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, ACM (New York, NY, USA, 2007), 617–626.

[12] Jarvis, J., Putze, F., Heger, D., and Schultz, T. Multimodal person independent recognition of workload related biosignal patterns. In *Proceedings of the 13th International Conference on Multimodal Interfaces*, ICMI '11, ACM (New York, NY, USA, 2011), 205–208.

[13] Kohlmorgen, J., Dornhege, G., Braun, M., Blankertz, B., Müller, K.-R., Curio, G., Hagemann, K., Bruns, A., Schrauf, M., and Kincses, W. Improving human performance in a real operating environment through real-time mental workload detection. In *Toward Brain-Computer Interfacing*. 2007, 409–422.

[14] Kothe, C., and Makeig, S. Estimation of task workload from EEG data: New and current tools and perspectives. In *Proceedings of the Engineering in Medicine and Biology Society* (2011), 6547–6551.

[15] Lei, S., and Rötting, M. Influence of task combination on EEG spectrum modulation for driver workload estimation. *Human Factors 53*, 2 (2011), 168–179.

[16] Murata, A. An attempt to evaluate mental workload using wavelet transform of EEG. *Human Factors 47*, 3 (2005), 498–508.

[17] Wang, Z., Hope, R. M., Wang, Z., Ji, Q., and Gray, W. D. Cross-subject workload classification with a hierarchical bayes model. *NeuroImage 59*, 1 (2012), 64–69.

[18] Wilson, G. F., Lambert, J. D., and Russell, C. A. Performance enhancement with real-time physiologically controlled adaptive aiding. *Proc. of the Human Factors and Ergonomics Society Annual Meeting 44*, 13 (2000), 61–64.

[19] Wilson, G. F., and Russell, C. A. Performance enhancement in an uninhabited air vehicle task using psychophysiologically determined adaptive aiding. *Human Factors 49*, 6 (2007), 1005–1018.