The Red-Line of Workload: Theory, Research, and Design

Conference Paper *in* Proceedings of the Human Factors and Ergonomics Society Annual Meeting · September 2008

DOI: 10.1177/154193120805201811

7 authors, including:

Rebecca A. Grier
Microsoft
38 PUBLICATIONS **1,427** CITATIONS
SEE PROFILE

Christopher D. Wickens
University of Illinois, Urbana-Champaign
49 PUBLICATIONS **2,731** CITATIONS
SEE PROFILE

David B Kaber
University of Florida
254 PUBLICATIONS **7,874** CITATIONS
SEE PROFILE

David L Strayer
University of Utah
226 PUBLICATIONS **14,054** CITATIONS
SEE PROFILE

The Red-Line of Workload: Theory, Research, and Design

Rebecca Grier, Christopher Wickens, David Kaber, David Strayer,
Deborah Boehm-Davis, J. Gregory Trafton, & Mark St. John

Multi-tasking is now ubiquitous component of our lives; despite the fact that we all can cite an incident where multi-tasking put us in a difficult situation. The reason so many of us do multi-task is that most of the time we are capable of effective dual task performance. Hart and Wickens (2008) have defined the point where one traverses safe and effective multi-tasking to dangerous and ineffective multi-tasking as the "red-line" of workload. In this panel, we will discuss this "red-line" of workload from the theoretical, empirical, and practical viewpoints. To that end, we first examine what theories of attention can help guide empiric search for this red line and where these theories must be expanded with further research. The greatest need is research that will allow human factors practitioners to identify the red line of workload before a system has been developed. One approach to achieving this research is to leverage the approach of industrial ergonomics, which has successfully defined physical workload limits by using data from safety incidents. Another avenue of research to be discussed is that which will lead to refinement of our theories and understanding of cognitive function to improve our ability to predict the red line. Next we move to the problem of evaluating systems to ensure that the red line of workload is not crossed. In particular, we will discuss the possibility of using task analysis, specifically, CPM-GOMS to predict if a system design will lead to excessive workload. Finally, we present two system design strategies for maintaining a cognitive workload that is below the red-line. The first of these is an adaptive automation using eye-tracking to reduce screen clutter when it appears workload has become so high an error may occur. The second design strategy presents four research based design principles for reducing workload to acceptable levels.

The Quest for the Red-Line of Workload.

Christopher D. Wickens

Modeling human multi-tasking performance continues to be a major challenge for human factors and cognitive engineering. Such modeling can be conceptualized as addressing three components:

1. When tasks are done concurrently, as in driving while conversing, we seek to understand properties that both allow this success and that modulate the degree of success (from perfect time sharing, to circumstances in which one or both tasks show a modest decrement). Multiple resource models (e.g., Wickens, 2008) characterize three influences on this concurrent processing: resource demand, resource structure, and the allocation policy between concurrently performed tasks, which determines which task suffers.
2. When tasks are performed sequentially, because of high demands, we seek "rules" or models of sequential processing, many based on queuing theory: what features determine the likelihood that ongoing tasks will be abandoned by interruptions (Trafton, 2008; Raby & Wickens, 1994) Can we predict how soon an abandoned task will be resumed? And its quality upon resumption?
3. Most critically, can we predict **at what level of demand**, the generally successful performance in (1) regresses to the sequential strategy in (2), in which one task or the other must be delayed? Answering this question of course, lies at the heart of a variety of **workload prediction** efforts**.** That is, can we objectively measure mental workload and, on such a metric, identify a "value" (or range of values") above which one or the other member of a task pair is "shed". Workload researchers have referred to this as the "red line" of workload (Hart & Wickens, 2008).

Defining such a red line, while of important theoretical interest, is also vital for two different human factors applications. First, those involved in certifying systems (e.g., new aircraft, or ATC systems) would like to know that "workload is acceptable" on such systems (e.g., that an air traffic controller can handle no more than N aircraft on a sector). The concept of "acceptable" then implicitly acknowledges some "red-line" of acceptability, along a workload scale. Second, developers of complex multi-task performance models such as IMPRINT (Laughery et al, 2006), or MIDAS (Gore & Jarvis, 2005) seek objective, empirically based criteria for triggering "workload management strategies", related to task shedding or postponement. That is, a red line above which the computation of workload is assessed to be excessive; and sequential processing routines are invoked within the model.

Of course, such a red line can always be drawn arbitrarily. But a non-arbitrary approach seeks what I refer to as a discontinuity or "knee" in a demand-performance curve, whereby performance loss either begins, or escalates, as the particular level of resource demand (workload) is exceded. For single tasks, such "knees" have been reported in various dimensions: e.g., a working memory limit of 4-5 chunks (Card, Moran & Newell, 1983), a relational complexity limit of 3 (the number of related components that must be held in working memory; Halford et al, 2005), or a speed of simple decision limit of 2.5/sec (Debecker and Desmedt 1970). However such metrics will be task and measure specific, and hence unsuited for heterogeneous dual task processing, where such single task metrics need to be combined to produce an overall multi-task workload metric, not specified by the parameters of a particular task. With the goal of defining more general metrics, some investigators have proposed particular levels along subjective

scales (e.g., a SWAT rating of 40; Reid & Colle, 1988; or an IMPRINT workload computation rating of 60; Mitchell et al, 2003). Such efforts, while on the correct path, lack empirical validation of the association of these values with a knee or discontinuity in multitask performance. Some alternative approaches will be discussed.

What can Industrial Ergonomics and Adaptive Automation Research Tell us about Defining a "Red-line" for Cognitive Workload?

David Kaber & Prithima Mosaly

Hart and Wickens (2008) proposed the concept of a "red-line" of cognitive workload for predicting failures in multitasking performance. They proposed the red-line could also be used to determine the acceptability of systems design, and provide a basis for triggering workload management aids in adaptive systems. Wickens (this panel) said the red-line could be objectively defined by an observed slowing in the rate of performance associated with increases in overall task demands.

In general, the red line concept can be related to historical physical workload limits defined based on industrial ergonomics research, such as the NIOSH lifting equations (Waters et al., 1993) or Job Severity Index (Ayoub et al., 1983). One of the main criticisms of these limits was their focus on specific ergonomic criteria (e.g., biomechanical) and de-emphasis of others (physiological, psychophysical), when each criterion may be optimized under different task conditions. This criticism also holds true for indices of cognitive state (e.g., NASA-TLX, SAGAT, SWAT). These indices have been validated for application across domains, but may be more or less sensitive to specific task factors. Consequently, they fail to indicate "red-line", excessive workload, for different task conditions.

In response to this criticism, industrial ergonomics research (Kim, 1990) attempted to identify a region of physical loading, bounded by the intersection of trends on various physical ergonomic criteria, to represent a set of values for which performance problems and injury might occur. A similar approach may be defined for cognitive tasks in which a red-line of cognitive workload is considered along with other red-lines based on various cognitive constructs, such as situation awareness (SA), to form a 'red region' of hazardous cognitive states. (Wickens has also spoken of a "red zone" of cognitive workload (personal communication) to identify the potential for multi-tasking performance failures.)

Our prior research on adaptive automation in complex systems for operator workload and SA management in multitasking (e.g., Endsley & Kaber, 1999; Kaber & Endsley, 2004) suggests that the concept of a red region of hazardous cognitive states, based on multiple cognitive criteria, may be necessary to provide an effective basis for design or task aiding. In support of this, we have observed differential effects of types of automation and strategies to dynamic function allocations on cognitive load, SA and performance. Nevertheless, there are certain commonalities across the results of these studies regarding the types of automation that may be generally poor for supporting human information processing (i.e., those that push us into the red region). Whatsmore, we have found the requirement for human performance of specific types of information processing to be a critical determinant of human-machine systems success or failure. Therefore, as Wickens (this panel) has suggested, the cognitive workload red-line may be dependent upon the type of task or function in which an operator is engaged. Given potential tradeoffs in operator workload and SA based on automated task aiding, it seems all the more important to define a red region of hazardous cognitive states for interface and control design. Automation design recommendations aimed at addressing a cognitive red-line should also consider the type of change in human information processing as well as differences in cognitive responses. These needs may also have relevance to contemporary augmented cognition research, which has its origins in adaptive automation.

There is also a pressing need not only for empirical validation of specific workload metrics in terms of performance changes, but for the red-line to be supported by real-world incident rate data. Consider the scrutiny of the scientific basis for the proposed OSHA Ergonomics Rule for the prevention of work-related musculoskeletal disorders, and the ultimate reversal of the rule in 2000. With respect to the development of cognitive systems design guidelines, standards or even regulations for various domains, demonstrating a decline in operator performance with increasing cognitive demands under multi-tasking conditions through controlled experiments may not be enough for general acceptance of the red-line criteria in practice. To implement any red line of workload or region of hazardous cognitive states as a basis for design practice, the relation of cognitive workload and SA research to the reality of safety and performance in real-world complex systems will need to be made crystal clear.

In summary, the new concept of cognitive redlining may learn from the experience of industrial ergonomics and adaptive automation research in two ways: (1) there is a need to consider multiple cognitive criteria in defining complex system design constraints, such as a maximum permissible cognitive load (MPCL) or a cognitive-support action limit (CSAL); and (2) there is a need for more rigorous field research (investigation of "cognition in the wild") for validation of cognitive red lines, possibly extending beyond what the human factors community has already engaged in, specifically correlating safety incident rates with values of the indices used to define red lines.

Towards a Red Line Metric for Multitasking in the Automobile

David Strayer

Establishing a theoretically driven and empirically validated red line metric of workload for multitasking activities in the automobile would be of considerable importance in the design and regulation of emerging technologies. However, several factors complicate the development of such a metric. For example, some dual-task combinations, such as engaging in a conversation over a cell phone or with a passenger in the vehicle, might appear at first glance to place a similar load on the driver. In fact, passenger conversations are qualitatively

less distracting than cell phone conversations because the structure of the conversation is dynamically altered by the dyad in these two situations. Another complication is that some dual-task combinations may produce a clear pattern of task switching (e.g., driving and text messaging) whereas other dual-task combinations may involve time-sharing (e.g., driving and conversing on a cell phone) or a combination of task switching and time-sharing. Moreover, despite the fact that several states have regulations that prohibit motorists from using hand-held cell phones but permiting the use of hands-free cell phones, the pattern of dual-task interference is identical for these two modes of cell phone use. This indicates that the locus of interference is due to cognitive interference rather than to manual interference (i.e., the interference is because the cell phone driver is not attending to the road causing a form of inattention blindness whereby they may fail to see up to half of the information in the driving environment). The pattern can become even more complicated in that a cell phone conversation that impairs one driver can be recorded and then played back to a different driver. Only the driver engaging in the initial conversation exhibits dual-task interference, suggesting that the generative components of speech are more disruptive than the portions of the conversation associated with comprehension. Similarly, listening to radio broadcasts and books on tape does not appear to impair driving performance. Data suggest that in some instances the verbal task of speech generation may elicit visual imagery that conflicts with the spatial codes required for the safe operation of a motor vehicle (interestingly, the nature of passenger and driver conversation appears to minimize this code conflict). Patterns of dual-task interference are also modulated by age and expertise. Novice drivers suffer greater dual-task interference than college students and in absolute terms drivers over 65 years of age exhibit the greatest costs of all (e.g., slowest braking reaction times) when conversing on a cell phone. Finally, important individual differences are emerging that indicate that not everyone suffers from the dual-task combination of driving and conversing on a cell phone. Approximately 3% of drivers who have high levels of executive control, as measured using the operation span task, actually perform better at both the driving and cell phone surrogate when performed in dual-task combination than when each is performed in single-task conditions. We are currently testing to see if these "super-taskers" are recruiting different portions of dorsal lateral prefrontal cortex to perform these dual-task activities. Refining our theoretical understanding of executive function and cognitive control in multitasking situations is critical to the establishment of a red line metric for workload in the automobile.

Can Cognitive Task Analyses Inform Workload Estimates?

Deborah A. Boehm-Davis

Cognitive task analysis techniques, which have been successful in predicting the time required for people to execute specific tasks using complex systems, also have the potential to aid in predicting workload, and in identifying the illusive "red line." This presentation will explore the extent to which CPM (Cognitive-Perceptual-Motor or Critical Path Method) GOMS (Goals-Operators-Methods-Selection Rules) can be useful in predicting workload (in general) and redline (in particular).

CPM-GOMS is a task analysis technique (Gray, John, & Atwood, 1993) that breaks tasks down to their elementary cognitive, perceptual, and motor operators. Actions taken within each "channel" (e.g. perceptual) are represented sequentially. Dependencies across channels are represented, but activities that can occur in parallel can do so. This technique yields a "critical path" that represents the shortest time from the beginning to the end of the task for expert users performing without error. When the execution of an operator (e.g. pressing a button) must wait for another operator (e.g., noticing that a light is lit), "slack" time is created on a channel. That is, there is time when the user is not using the resources from that channel and where that channel's resources could be used for other appropriate activities. A measure of resources available (an inverse of workload) might be leveraged from summing the amount of slack time available across all channels being used, subtracting it from the total time required to execute the critical path, and dividing by the total time:

$$\left(1 - \frac{CriticalPathTime - SlackTime}{CriticalPathTime}\right).$$

This "available" time might provide an indication of how many resources are available for additional task demands or additional tasks. Such a metric would suggest that decreases in the amount of resources available are indicative of increases in workload to a point of saturation where the "red line" is reached.

However, such a metric suffers from several flaws. Although one might argue that at some percentage, it becomes impossible to have sufficient time to allow for either the interpretation of unexpected events or indeed, the execution of any other secondary task, this metric does not provide an indication of where the "red line" would fall. Second, it hinges on assumptions that may not be reasonable in this context. Specifically, the times used in this analysis technique assume both expert performance and no errors, both of which may not be appropriate in situations where tasks (or specific combinations of tasks) are performed rarely. Second, the parameters used to estimate time for each step in task (e.g., perception) are fixed and assume routine behavior; that is, the technique does not allow time for interpretation or decision-making Third, the technique also assumes that users know what to look for; to the extent that interfaces either label an action in a non-obvious way, the parameters used will fail to accurately predict performance. Finally, the technique assumes that users will know how best to use the slack time available to perform the secondary task.

Nonetheless, using the amount of slack time available based on a CPM-GOMS analysis does provide a scaled metric of potentially-available resources. Although this may be a relatively crude measure, it is more fine-grained than other techniques currently available for predicting workload. CPM-GOMS is a well-established technique and this approach may provide platform for further investigation and validation.

Perceptually Predicting Post-Completion Errors

J. Gregory Trafton & Raj M. Ratwani

Hart and Wickens (2008) and Wickens (this panel) proposed the concept of a red-line of workload: accurately determining workload in order to know that the level of workload is acceptable for a given task or to provide task facilitation when workload gets too high. One component of high workload is a decrease in performance and an increase in number of errors. Our focus is on predicting when someone is likely to make an error so that we can provide task facilitation in order to reduce the chance of errors.

Our proposed solution relies on two insights: First, researchers should rely on theory to drive the search for applied answers (Trafton & Altmann, under review; Wickens, 1974-2008). Second, that by using measures of online perception, it should be possible to predict certain classes of errors.

Our theoretical perspective comes from the Memory for Goals theory (Altmann & Trafton, 2002, 2007), which suggests that goals are forgotten due to a combination of decay and lack of priming from environmental cues. If we assume that many errors are due to forgotten goals, we can use that information to predict when people would be likely to make an error. In this work, we focus on post-completion errors, which are associated with an action that is required after the main goal of the task has been completed, like leaving the original in the copier after making your copies. However, since Memory for Goals is primarily an associative memory model and it is notoriously difficult to make online predictions about what a particular memory element's activation is at a particular point in time, it is probably best not to rely solely on such a memory model until the models become better at their predictive ability. However, we can use perception – eye-movements in particular – as a window into the mind.

We have taken these two theoretical predictions – errors are caused by decay (operationalized as time) and environmental cues (operationalized as whether the correct step was looked at), examined data on error episodes, and created a logistic regression equation to predict errors. We found that, at least for some classes of errors, our theoretically derived measures are able to predict almost 90% of the errors that people make before they make them. This theoretical model has been cross-validated.

From an applied perspective, we have created an online system that tracks people's eye-movements and presents a cue only when the equation predicts that there is a 75% of making an error. This approach reduces interface clutter and provides a "just-in-time" method of predicting when an error will occur.

One of our long-term goals is to cross-validate the empirically derived parameters on a different task. If our approach is correct, we may have identified the red-line for predicting post-completion errors.

Four Design Principles for Maintaining and Recovering Situation Awareness

Mark St. John & Harvey S. Smallman

Multi-tasking and the interruptions inherent to them pose a challenge for maintaining good performance in many tasks. For dynamic operational tasks, such as airspace monitoring and civil emergency operations, the situation changes over time and users must detect and understand those changes both during real-time monitoring and following interruptions in order to maintain situation awareness and task performance.

Poor change detection ability makes this maintenance difficult enough while monitoring a situation display uninterrupted, and multi-tasking and interruptions increase the difficulty dramatically. Yet little research has addressed the need for better interface tools to help users detect and interpret changes either to maintain situation awareness or recover it following interruptions.

To address this need, we followed the approach taken by Wickens and Carswell (1995) and their proximity-compatibility principle to identify design principles that generalize across particular tasks and artifacts. To this end, we very briefly describe results from experiments on two distinct dynamic tasks, air warfare and a team collaboration task, and derive four principles for the design of effective interruption recovery tools. Better designs may lower workload, increase productivity, and reduce manning.

The principles are 1) augment users' natural change detection ability with automated change detection processes, 2) notify users of automatically detected changes in a relatively unobtrusive manner so that they are noticeable but minimally distracting from on-going tasks, 3) provide summary descriptions of each significant change to allow users to scan and prioritize the order in which changes are reviewed, 4) for busy, cluttered displays, make change information available only on demand by the user. This fourth principle, access on demand, raises complex issues regarding the trade-offs between display clutter and different modes of information access (e.g., Yeh & Wickens, 2001; Wickens, Sebok, Bagnall, Kamienski, & 2007).

Lastly, we contrast a number of interface designs in terms of the principles. Consideration of the four principles, as well as new ones to be identified, should facilitate the design of more effective tools to help users recover situation awareness for these important and high risk tasks.

References

Altmann, E. M. & Trafton, J. G. (2002). Memory for goals: An activation-based model. Cognitive Science. 26(1), 39-83.

Altmann, E. M. & Trafton, J. G. (2007). Timecourse of Recovery from Task Interruption: Data and a Model. *Psychonomics Bulletin and Review*, 14(6), 1079-1084.

Ayoub, M. M., Selan. J. L. & Liles, D. H. (1983). An ergonomics approach for the design of manual materials handling tasks. *Human Factors, 25*, 507-515.

Card, S. Moran,. T & Newell, A. (1986) The model humanprocessor. In K. Boff, L Kafman & J Thomas (EDS) Handbook of perceptuaionand human performance: vol 2: NY: John Wiley.

Debecker, J & Desmedt, R (1970) Maximum capacity for sequential one-bit auditorydecisions. *Journal of Experimental Psychology*. 83, 266-373.

Endsley, M. R. & Kaber D. B. (1999). Level of automation effects on performance, situation awareness and workload in a dynamic control task. *Ergonomics, 42*(3), 462-492.

Gore, B. F. & Jarvis, P. A. (2005) New integrated modeling capabilities: MIDAS' recent behavioral enhancements. (SAW-2005-01-2701) Warrandale, PA: SAE.

Gray, W. D., John, B. E., & Atwood, M. E. (1993). Project Ernestine: Validating GOMS for predicting and explaining real-world task performance. *Human-Computer Interaction, 8*(3), 237-309.

Halford, G.S. Baker, R., McCreddedn, J., & Bain, J.d. (2005). How many variables can humans process? *Psychological Science, 16,*70-76.

Hart, S, and Wickens, C.D. (2008) Mental Workload. In NASA Human Integration Design Handbook.

Kaber D. B. & Endsley, M. R. (2004). The effects of level of automation and adaptive automation on human performance, situation awareness and workload in a dynamic control task. *Theo. Issues in Ergo. Sci., 5*(2), 113-153.

Kim, H. K. (1990). Development of a model for combined ergonomic approaches in manual material handling tasks. Unpublished PhD dissertation. Texas Tech Univ.

Mitchell, D. Samms, C., Henthorn, T and Wojciechowski, J (2003). Trade study: a two versus three soldier crew for the mounted combat system. Army Research Lab technical report ARL-TR-3026.

Reid, G. B., & Colle, H. A. (1988) Critical SWAT values for predicting operator overload. *Proceedings of the Human Factors Society (HFES) 32nd Annual Meeting*. Santa Monica, CA: HFES, 1414-1418.

Trafton, J. G. & Altmann, E. M. (under review). A primer on how to do theoretically applied research: A case study of interruptions.

Waters, T.R., Putz-Anderson. V. et al.(1993). Revised NIOSH equation for the design and evaluation of manual lifting tasks. *Ergonomics, 36*, 749-776.

Laughery, K. R., LeBiere, C., & Archer, S. (2006). Modeling Human Performance in Complex Systems. In G. Salvendy (Ed.), Handbook of Human Factors and Ergonomics (3rd Ed.) (pp. 965-996). Hoboken, NJ: John Wiley & Sons.

Raby, M., & Wickens, C.D. (1994). Strategic workload management and decision biases in aviation. *International Journal of Aviation Psychology*, *4*(3), 211-240.

Trafton, G. (2008, in press) Dealing with Interruptions. In D. Boehm-Davis (Ed) *Reviews of Human Factors & Ergonomics. 3.*

Wickens, C.D. (2008 in press). Multiple Resources and mental workload. *Human Factors 50.*

Wickens, C. D. & Carswell, C.M. (1995). The proximity compatibility principle: Its psychological foundations and its relevance to display design. *Human Factors*, *37*, 473-494.

Wickens, C. D., Sebok, A., Bagnall, T., & Kamienski, J. (2007). Modeling of situation awareness supported by advanced flight deck displays. In *Proceedings of the HFES 51st Annual Meeting*. Santa Monica CA: HFES.

Yeh, M. & Wickens, C. D. (2001b). Attentional filtering in the design of electronic map displays: A comparison of color coding, intensity coding, and decluttering techniques. *Human Factors, 43,* 543-562.

Waters, T.R., Putz-Anderson. V., Garg, A. & Fine, L. J. (1993). Revised NIOSH equation for the design and evaluation of manual lifting tasks. *Ergonomics, 36*, 749-776.