Luca Longo
M. Chiara Leva (Eds.)

Communications in Computer and Information Science 1012

Human Mental Workload

Models and Applications

Second International Symposium, H-WORKLOAD 2018
Amsterdam, The Netherlands, September 20–21, 2018
Revised Selected Papers

Springer

Communications in Computer and Information Science 1012

Commenced Publication in 2007
Founding and Former Series Editors:
Phoebe Chen, Alfredo Cuzzocrea, Xiaoyong Du, Orhun Kara, Ting Liu,
Krishna M. Sivalingam, Dominik Ślęzak, and Xiaokang Yang

More information about this series at http://www.springer.com/series/7899

Luca Longo · M. Chiara Leva (Eds.)

Human Mental Workload

Models and Applications

Second International Symposium, H-WORKLOAD 2018
Amsterdam, The Netherlands, September 20–21, 2018
Revised Selected Papers

Springer

Editors
Luca Longo
Dublin Institute of Technology
Dublin, Ireland

M. Chiara Leva
Dublin Institute of Technology
Dublin, Ireland

ISSN 1865-0929 ISSN 1865-0937 (electronic)
Communications in Computer and Information Science
ISBN 978-3-030-14272-8 ISBN 978-3-030-14273-5 (eBook)
https://doi.org/10.1007/978-3-030-14273-5

Library of Congress Control Number: 2019932782

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

This book endeavors to stimulate and encourage discussion on mental workload, its measures, dimensions, models, applications and consequences. It is a topic that demands a multidisciplinary approach, spanning across human factors, computer science, psychology, neuroscience, statistics, and cognitive sciences. This book presents recent developments in the context of theoretical models of mental workload and practical applications.

The book and its central theme arose in the context of the Second International Symposium on Mental Workload, Models and Applications (H-WORKLOAD 2018), sponsored by the Netherlands Aerospace Centre (NLR) and supported by the Irish Ergonomics Society. It contains a revision of the best papers presented at the symposium and selected through a strict peer-review process. From the content of these research contributions, it is clear that mental workload is a very complex construct, thought to be multidimensional and multifaceted, with no clear and accepted definition. This is confirmed by the different modeling approaches that consider different factors and employ distinct strategies for their aggregation. Despite this uncertainty in modeling mental workload, it is clear that this construct is key for predicting human performance.

The nature and efforts required by the modern workplace is shifting toward more complex cognitive demands, as Professor Catherine Burns pointed out, and predicting the influence of new interactive technologies on human work is a critical capability for the practitioners of the future. The capacity to assess human mental workload is a key element in designing and implementing processes capable of monitoring interactions between automated systems and the humans destined to use them. Similarly, mental workload assessment is key for designing instructions and learning tools aligned to the limitations of the human mind. Unfortunately, mental workload measurement is not trivial. Some of the articles published in this book applied psychological subjective self-reporting measures, others made use of physiological or primary task measures, and some a combination of these. We believe the adoption of a multidimensional approach is fundamental to further understand the complex construct of mental workload and to fully grasp its nature. A number of research articles in this book have started focusing on the development of novel models of mental workload employing data-driven techniques, borrowed from machine learning as subfield of artificial intelligence, whose explanatory capacity over the topic is still to be explored. This last area is a field where traditional human factor approaches and novel data-driven modeling approaches can cross paths and perhaps trace a new fundamental research direction that should be further explored and promoted.

We wish to thank all the people who helped in the Organizing Committee for the Second International Symposium on Mental Workload, Models and Applications (H-WORKLOAD 2018). In particular the local chairs, Rolf Zon, Wendy Duivestein, Tanja Bos, and many more of the members of the Scientific Committee. We want to

also thank the main sponsors of the event, the Netherlands Aerospace Centre and the Irish Ergonomics Society, without which neither the conference nor the book would have been realized. A special thanks goes to the Dublin Institute of Technology as well as all the reviewers of the Program Committee who provided constructive feedback. A special thanks goes to the researchers and practitioners who submitted their work and committed to attending the event and turning it into an opportunity to meet and share our experiences in this fascinating topic.

January 2019

Luca Longo
M. Chiara Leva

Organization

Organizing Committee

General Chairs, Editors, and Program Chairs

Luca Longo	Dublin Institute of Technology, Ireland
M. Chiara Leva	Dublin Institute of Technology, Ireland

Local Chairs

Rolz Zon	Netherlands Aerospace Centre, The Netherlands
Tanja Bos	Netherlands Aerospace Centre, The Netherlands

Program Committee

Aghajani Haleh	Baylor Scott & White Research Institute, USA
Balfe Nora	Trinity College Dublin, Ireland
Barge Roland	Rolls-Royce, UK
Bellini Emanuele	LOGOS Ricerca & Innovazione, Florence, Italy
Bos Tanja	Netherlands Aerospace Centre, The Netherlands
Brookhuis Karel	Rijks Universiteit Groningen, The Netherlands
Byrne Aidan	ABMU Health Board, UK
Cahill Joan	Trinity College Dublin, Ireland
Cain Brad	C3 Human Factors, Canada
Castor Martin	GEISTT, Sweden
Cerrato Loredana	EITDigital, Sweden
Charles Rebecca	Cranfield University, UK
Costa Antonio Pedro	Aveiro University, Portugal
Edwards Tamsyn	NASA Ames/San Jose State University, USA
Genga Glauco	Institute of Aerospace Medicine, Italy
Galicic Matjaz	Arcadis, UK
Golightly David	HFRG, University of Nottingham, UK
Greiner Birgit	University College Cork, Ireland
Hines Andrew	University College Dublin, Ireland
Kane Bridget	Karlstad University, Sweden
Kenvyn Fiona	Metro Trains Melbourne, Australia
Konstandinidou Myrto	NCSR Demokritos, Greece
Maij Anneloes	Netherlands Aerospace Centre, The Netherlands
Massaiu Salvatore	IFE, Norway
Mijovic Pavle	mBrainTrain, Serbia
Mills Ann	Railway Safety Standards Board, UK
O'Sullivan Leonard	University Limerick, Ireland

Orru Giuliano	Dublin Institute of Technology, Ireland
Pediconi Maria	University of Urbino, Italy
Pickup Laura	University of Exeter, UK
Raduntz Thea	Federal Institute for Occupational Safety and Health, Germany
Ratkai Melinda	University of Malaga, Spain
Raufi Bujar	South East European University, Macedonia
Rizzo Lucas	Dublin Institute of Technology, Ireland
Rocha Mariana	Dublin Institute of Technology, Ireland
Roelen Alfred	Netherlands Aerospace Centre, The Netherlands
Smith K. Tara	Human Factors Engineering Solutions Ltd., UK
Teixeira Carlos	Dublin City College, Ireland
Waldmann Thomas	University of Limerick, Ireland
Wilkins Maurice	Yokogawa, UK
Young Mark	Loughborough, UK
Zon Rolf	Netherlands Aerospace Centre, The Netherlands
De Waard Dick	University of Groningen, The Netherlands
Moustafa Karim	Dublin Institute of Technology, Ireland
Van Dijk Henk	Netherlands Aerospace Centre, The Netherlands

Contents

About the Editors

Dr. Luca Longo is currently Assistant Professor at the Dublin Institute of Technology, Technological University Dublin, where he is a member of the Applied Intelligence Research Centre and ADAPT, the Global Center of Excellence for Digital Content and Media Innovation. His core theoretical research interests are in artificial intelligence, specifically in automated reasoning and machine learning. He also performs applied research in mental workload modeling. He is author of 45+ peer-reviewed articles appeared in conference proceedings, book chapters, and journals in various theoretical and applied computer science fields. Luca was awarded the National Teaching Hero in 2016, by the National Forum for the Enhancement of Teaching and Learning in Higher Education.

Dr. Maria Chiara Leva is Lecturer at the Dublin Institute of Technology in the College of Environmental Health. She is also a visiting research fellow in the Centre for Innovative Human Systems at Trinity College Dublin. Her area of expertise is human factors and safety management systems. Chiara holds a PhD in human factors conferred by the Polytechnic of Milan, Department of Industrial Engineering. Her PhD focused on human and organizational factors in safety-critical system in the transport sector. She is the co-chair of the Human Factors and Human Reliability Committee for the European Safety and Reliability Association and has been working in ergonomics and risk assessment as a consultant since 2008. She has more than 56 publications on human factors, operational risk assessment, and safety management in science and engineering journals.

Models

Understanding, Supporting, and Redesigning Cognitive Work

Catherine M. Burns(✉)

Systems Design Engineering, University of Waterloo, Waterloo, Canada
catherine.burns@uwaterloo.ca

Abstract. Cognitive work analysis (CWA) is a framework that has been used in many settings to describe various aspects of work. This paper outlines how CWA can be used to understand work and mental workload. The work domain, control task, and strategies analysis can be useful to understand the nature of work, work allocation and mental workload. Finally, the prediction of work patterns is discussed. Predicting the influence of new technologies on human work is a critical capability for the human factors practitioners of the future.

Keywords: Cognitive Work Analysis · Mental workload · Task analysis · Function allocation

1 Introduction

As Wickens [23] introduced at H-WORKLOAD last year, studies of mental workload began in the 1960’s and have accelerated through the decades. Mental workload is certainly not a mature science, and I expect that it will always be an exciting area of study. Exciting new methods of being able to compute and predict mental workload are emerging [12–15, 18]. While it is true that studies of mental workload measurement techniques such as the NASA-TLX have reached maturity, there is a significant new challenge on the horizon. That challenge is the prediction of the impact on humans of new technologies. New forms of automation and artificial intelligence have the potential to impact human work at a cognitive level we have not seen before (e.g. [5, 6, 11, 20]). Being able to predict the impact of these new technologies, so that benefits and risks can be anticipated, will be an important contribution to society for those who practice human factors and ergonomics.

Our last significant technological revolution was the arrival of the internet, and it dramatically changed the nature of human work and communication around the world. In 1982, a New York Times article made reasonable predictions that the emergence of the internet would improve communications, allow individuals to create their own content, and blur lines between home and work by allowing more work at home. Other implications such as advanced globalization, economic shifts and the development of social and political hacking through things like fake news were not predicted. Had we been able to predict some of these effects, safeguards and more effective responses could have been developed. There are some differences though in these two technological revolutions. The internet largely was about connectivity, and improved connectivity was perhaps not in the domain of human factors professionals. However, the reallocation of human decision

L. Longo and M. C. Leva (Eds.): H-WORKLOAD 2018, CCIS 1012, pp. 3–12, 2019.
https://doi.org/10.1007/978-3-030-14273-5_1

making to automation and artificial intelligence, and the resultant supervisory control problems land squarely within the expertise of our community. I am not arguing that human factors practitioners become futurists, but we do have an important role to play in the careful prediction of how new technologies will influence how people work and live.

Suppose a client asks you whether bringing in a new artificial intelligence system will allow her physicians to make better decisions. The core question of whether decision making will better with the new tool can be answered (relatively) easily. The tool can be analyzed and its accuracy compared to expected human performance. If funding allows, we can run a study to compare the performance of physicians using the tool, against the performance of those physicians who are not using the tool. These are all good questions, and solid approaches exist for answering these questions. More challenging questions though are:

- Will I still need as many physicians?
- Should I still pay the physicians the same amount?
- How will this tool influence their mental workload?
- Will I be expecting my physicians to analyze more cases per day?
- Does the tool take away their decision-making latitude, making them mechanical tool users where before they had rich and empowered positions?
- Is the tool telling them enough of how it made its decisions that they can justify their case to a concerned family member, or defend themselves in a negligence case?

These are the challenging questions of how people work with technology today. I firmly believe that we have an important role in being a part of that conversation. Our field has the tools and methods that can begin to identify these kinds of issues. Our approaches are systematic and scientific and can move us away from the realm of speculation. In this paper, I will discuss how one of our tools, Cognitive Work Analysis (CWA), can be used for mental workload analysis. From this abstraction, I believe that most of our work analysis approaches can be used to answer mental workload prediction questions. Although I am using CWA as my base for discussion, there are many methods within our human factors toolbox that can be used to model and understand mental workload (e.g. [13, 23]). I want to encourage us to think of the decision latitude we give our users through design, how we can make decision making more effective, and how we can predict the adaptation of work.

Cognitive Work Analysis (CWA) is well known as a framework for analyzing work [16, 17, 22]. CWA takes a multi-faceted view of human work, looking at work through five different lenses: domain constraints (work domain analysis), tasks and information processing (control task analysis), strategies, social, organizational constraints, and worker competencies. Many times, CWA is used in environments that are, by their very nature, complex and CWA is a useful method for unpacking that complexity and identifying the areas where users need more support for their work. In particular, the first three phases of CWA, work domain analysis, control task analysis, and strategy analysis reveal important aspects of human work and how to design healthy and effective work.

CWA is inherently focused on the analysis of work, and yet, CWA is not used explicitly for mental workload analysis by most CWA practitioners. In many ways, this is something of a lost opportunity. CWA can help to investigate work and to understand when work is complex, when the mental workload is high, or when work could be supported with technology. In this paper, I will step through the first three phases of CWA, identifying the mental workload implications implied by the models.

2 Work Domain Analysis: Building Decision Latitude

Work domain analysis (WDA) is known for its functional description of the work environment. In many ways, WDA is the differentiating analysis of CWA, providing a different view of the work environment than most other analyses provide. The WDA uses the abstraction hierarchy [17, 22] as its most common model, which looks at the world the person must operate in, the purposes of that system and how that system operates. The abstraction hierarchy is grounded on the assumption that an operator will be working towards a successfully functioning work system, and responding within the constraints of that system to achieve the required purposes. For a human factors practitioner, conducting a WDA brings the practitioner into the world of the operator, helps to show areas of complexity for that operator, makes the goals of the operator and their action possibilities apparent to the analyst. A WDA does not specify actions, or user interface technologies, or even ideal approaches. A WDA is not a human-centered analysis; it is an objective analytical re-engineering of the work environment to bring a greater understanding of the needs of workers in that environment. The human-centered, work-centered or human factors oriented advantages of WDA come from how the analysis is used, not from the analysis itself.

The abstraction hierarchy has implications for understanding mental workload. First, expert workers understand the components and constraints of their environment with more clarity than novice workers [2]. At first, this additional knowledge, of being aware of all the various purposes, components, processes and action possibilities, might seem like higher mental workload. However, with expertise being aware of purposes and action possibilities makes that worker a better decision maker. The awareness that various solutions are possible creates decision latitude for that worker through competency (Fig. 1). Work with higher levels of decision latitude is known to create more empowering and healthier work [8]. The work may still be stressful, and the mental workload may still be high, but the nature of that mental workload is more rewarding.

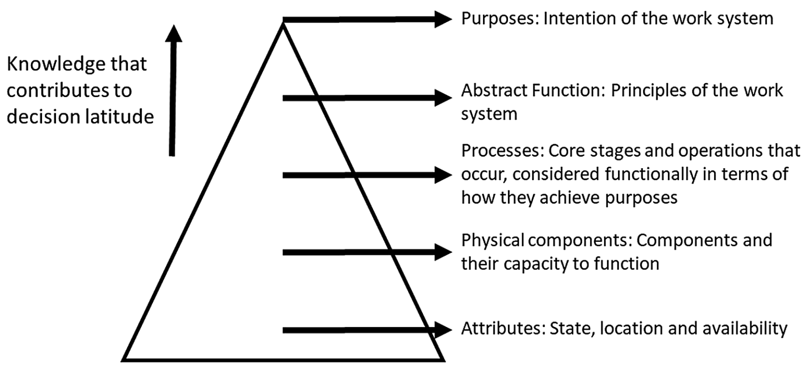


Fig. 1. Understanding functional relationships and purposes leads to more decision latitude for decision makers.

The corollary of this implication is that when we design better displays for workers, showing them more of the intention of the work system, and their capabilities to react to situations, we change the nature of their work. We know from research that we create workers capable of better diagnoses and better able to mitigate situations. We can develop workers with a greater sense of competency and empowerment to take successful action in their work environments. Figure 2 illustrates different intention and action distributions, described along the structure of the abstraction hierarchy.

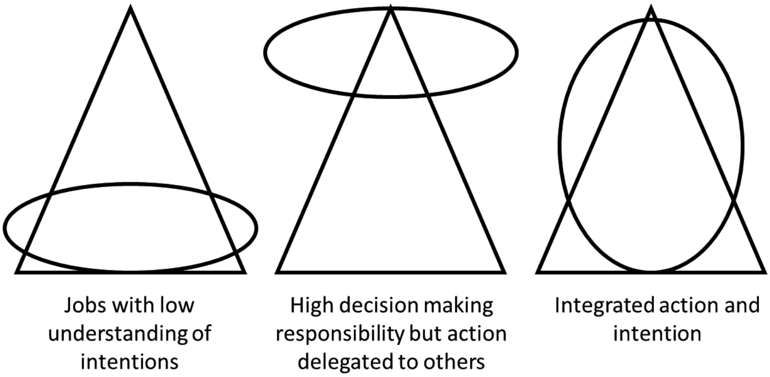


Fig. 2. Different role allocations with different mixes of intention understanding and capability for action.

Often the complexity of work requires the distribution of work through functional stratification. In several projects, we have developed role allocation maps along the abstraction hierarchy, and patterns like those in Fig. 2 have been noticed. Examples include Ashoori [1] which looked at roles across a team in a labor and delivery unit of a hospital and Hajdukiewicz [7] who looked at role division during surgery.

Not every role can have integration between the intentions and actions of the work system, particularly in larger work contexts where the distribution of work is required. However, in cases where workers must either focus on running processes and executing actions (the left case in Fig. 2), they will benefit from efforts to keep them informed of the intentions and goals of the system. Similarly, workers who must work as decision makers and managers will benefit from information on how other teams are making progress in executing plans and actions. Some of these ideas were presented in [4].

3 Control Task Analysis: Looking for Cognitive Efficiencies

The Control Task Analysis (ConTA), and its Decision Ladder (DL) is the phase of CWA that is used most often in discussions of work modification. The DL is a template of human information processing, and the intention of the analysis is to understand how tasks are triggering various processing steps. The ConTA presents alternative behaviors, with knowledge-based behavior and cognitively costly analytical decision-making at its top, and more efficient and less effortful heuristic and rule-based processing at the bottom of the DL (Fig. 3). The analysis intends to understand when cognitively effortful work is occurring and whether there is an opportunity to replace that work with approaches that are less effortful. The ConTA also models expertise and acknowledges that less experienced workers may spend more time in the more effortful processing. Experienced workers facing new situations will also revert to these processes at times when their experience cannot provide the heuristics for more efficient decision making.

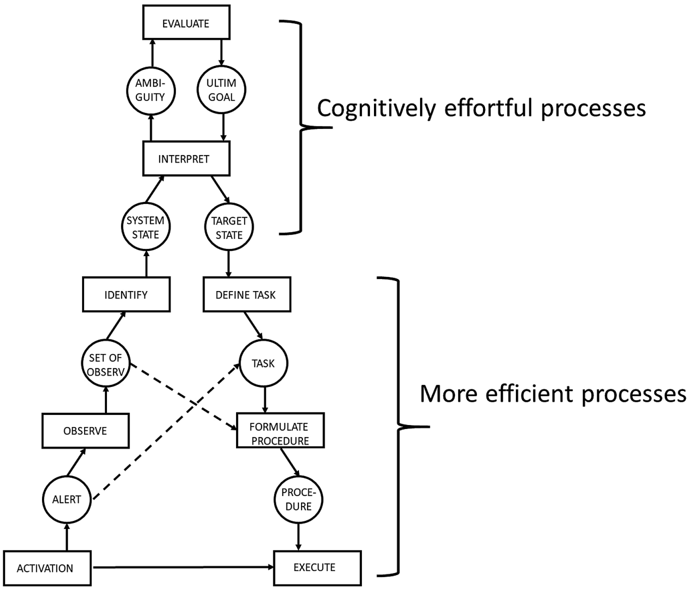


Fig. 3. The decision ladder can identify cognitively effortful processes.

A second application of the DL can be to identify technologies that are adding unnecessary mental workload. The DL can show where information is missing, and work processes become less efficient. We developed The DL below (Fig. 4) by observing pharmacists as they processed prescriptions and checked for medication issues [9]. The DL helped to identify that information such as allergy information or interactions was often not shared with the pharmacist. Because this information was missing, the pharmacist would need to go through an extra problem-solving process to identify these issues, possibly communicating with the patient or their physician. Including this information with the prescription or the medication record would have allowed these pharmacists to work more efficiently. A small change in information could have generated a significant improvement in mental workload.

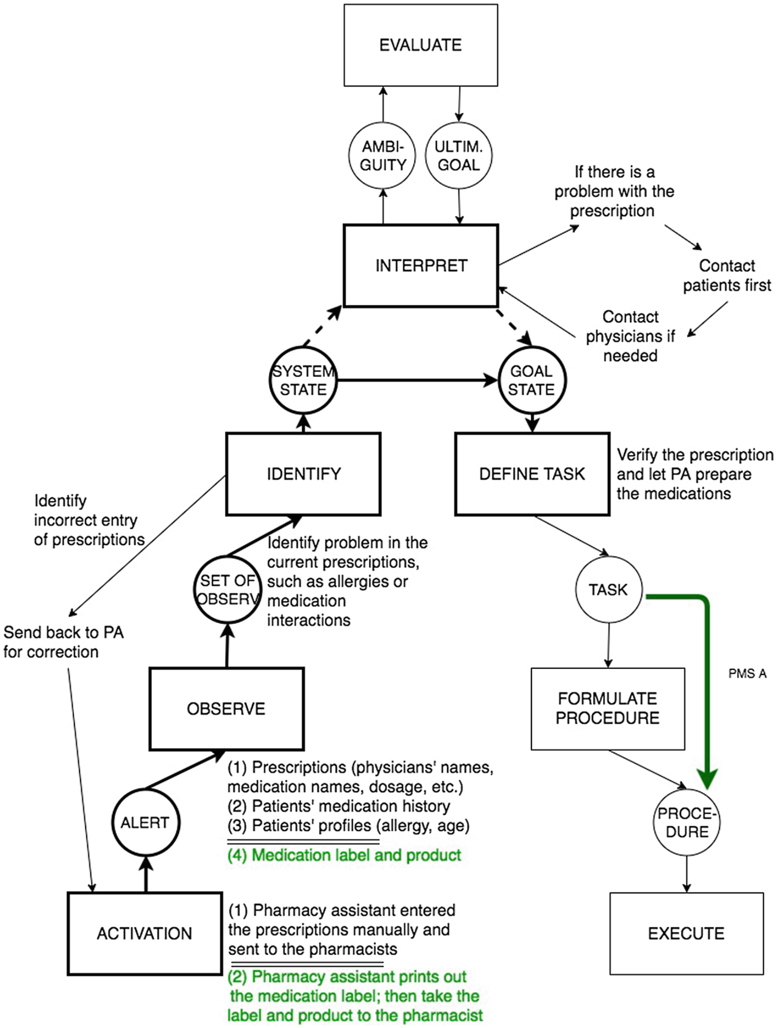


Fig. 4. Identifying technology that increases mental workload in the decision ladder.

4 Strategies Analysis: Understanding How Work Adapts

The Strategies Analysis (SA) looks at the different ways people can solve the same problem. Unlike the ConTA, the focus is not on information processing or decision-making. In this case, understanding the variety of pathways and what triggers put people on those paths is important. Most often, the triggers for different strategic paths are either particular contexts, the experience of the user, or different mental workload levels (Fig. 5). We know well that under high mental workload users will take different approaches, either shedding tasks or building workarounds. The SA can in this way be used to investigate high mental workload situations and to ask users how they adapt to these situations.

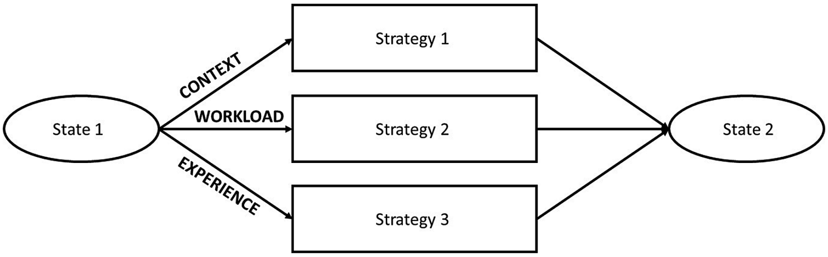
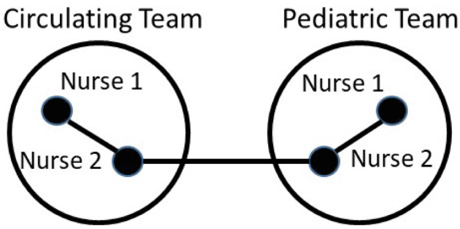
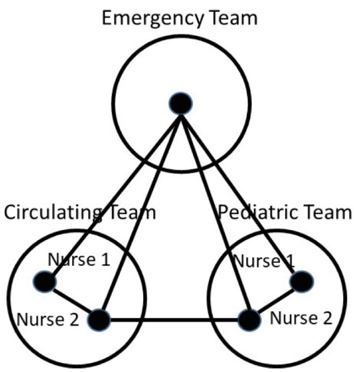


Fig. 5. Context, mental workload, and experience determine the strategies that are chosen.

Some of the adaptation strategies we have seen in other work has been adjusting mental workload patterns or reducing communication. In some cases, team structures have shifted to handle higher mental workload conditions. In the figure below (Fig. 6) we show strategic shifts in team composition to handle the additional mental workload of an emergency delivery. The normal delivery team structure is at the top; the emergency team



Strategy 1: Team configuration under normal conditions.



Strategy 2: Team configuration under high mental workload and critical conditions, centralized control.

Fig. 6. Different strategies for different mental workload situations, normal (*on top*) or emergency (*lower figure*) [1].

structure is below. The team has increased in size to handle the mental workload, but the nature of the work has also changed. In the emergency, the work is more challenging, and the emergency team has taken tighter control of the team structure [1].

5 Mental Workload Prediction

While analyzing existing work patterns can help in redesign, we are often asked to anticipate the effects of new technology in the workplace. The same methods can be used prospectively to predict where new technologies may add or reduce work, require new function allocations, or require additional information support. There are several specific ways to predict work.

- **Comparison of similar environments.** Models of similar work environments can be compared, and the differences in one work environment used to understand the other environment. Examples of these comparisons are [3, 21]. To make these comparisons, the two environments must share some functional similarity.
- **Analysis of the "cut points" in models.** Whenever work must be distributed over people, over function, or over process, "cut points" in work occur. Figure 2, for example, showed two ways that work could be distributed: separating decision making from operations. Whenever these separations occur, you can predict that workers will need support to understand the perspective and activities from the other work zones around them. These divisions can show points where situation awareness can be lost and where workers can become isolated in their work processes.
- **Analysis of the functionality and behavior of the technology.** Particularly when technologies provide assistance by covering some of the human work, it becomes critical to understand how that technology will interact. The decision ladder can be used to develop detailed and sensitive function allocation models, that show what roles the technology is expected to take [10]. The human work should be studied at the transition points, where the human moves work to the technology or takes work over from the technology. (Automated vehicle handovers are a classic example of this problem).

6 Conclusion

The understanding of human work continues to be an important and exciting field, even years after the origins of human factors and ergonomics. As rich and complex technologies are added to the workplace, it becomes critical to understand how people will interact with these technologies. New technologies have the potential to replace human work, but this needs to be done in a collaborative way, where humans and technologies are partners. This partnership means providing the communication to support smooth work transitions and continuing to position humans in roles where they have decision latitude as well as good situation awareness of actions that are occurring under their supervision.

Mental workload will remain a critical issue for the future, in particular, as humans move to more supervisory roles. The mental workload of working with technologies such as automation, artificial intelligence, and more complex systems, is still only loosely understood. More research is needed that explores the mental workload of working with these new technologies. Further, there is a need for a more finely understood concept of mental workload. There is great promise emerging in the development of new mental workload measurement techniques, from physiological measurements and neuroergonomic viewpoints to machine learning models. This conference, and the papers included in this issue, contribute invaluably to these new directions.

References

1. Ashoori, M., Burns, C.M., Momtahan, K., d'Entremont, B.: Using team cognitive work analysis to reveal healthcare team interactions in a labor and delivery unit. Ergonomics **57**, 973–986 (2014). https://doi.org/10.1080/00140139.2014.909949
2. Burns, C.M.: CWA: models of expertise. In: Ward, P., Schraagen, J.M., Gore, J., Roth, E. (eds.) Oxford Handbook of Expertise: Research and Application. Oxford University Press (in press). https://doi.org/10.1093/oxfordhb/9780188795772.001.0001
3. Burns, C.M., Bisantz, A.M., Roth, E.M.: Lessons from a comparison of work models: representational choices and their implications. Hum. Fact. **46**, 711–727 (2004). https://doi.org/10.1518/hfes.46.4.711.56810
4. Burns, C.M., Torenvliet, G., Scott, S., Chalmers, B.: Work domain analysis for establishing collaborative work requirements. In: Proceedings of the 53rd Annual Meeting of the Human Factors and Ergonomics Society, pp. 314–318. HFES, Santa Monica (2009). https://doi.org/10.1177/154193120905300432
5. Cahill, J., et al.: Adaptive automation and the third pilot: managing teamwork and workload in an airline cockpit. In: Longo, L., Leva, M.C. (eds.) H-WORKLOAD 2017. CCIS, vol. 726, pp. 161–173. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61061-0_10
6. Edwards, T., Martin, L., Bienert, N., Mercer, J.: The relationship between workload and performance in air traffic control: exploring the influence of levels of automation and variation in task demand. In: Longo, L., Leva, M.C. (eds.) H-WORKLOAD 2017. CCIS, vol. 726, pp. 120–139. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61061-0_8
7. Hajdukiewicz, J.R., Vicente, K.J., Doyle, D.J., Milgram, P., Burns, C.M.: Modeling a medical environment: an ontology for integrated medical informatics design. Int. J. Med. Inform. **62**, 79–99 (2001). https://doi.org/10.1016/S1386-5056(01)00128-9
8. Karasek, R.A.: Job demands, job decision latitude, and mental: strain: implications for job redesign. Admin. Sci. Quart. **24**, 285–308 (1979). https://doi.org/10.2307/2392498
9. Kerestegliocl u, D., Burns, C.M., Grindrod, K.: Building bridges between physicians and pharmacists: an interprofessional approach to electronic health record requirements. In: HFES 2016 International Symposium on Human Factors and Ergonomics: Shaping the Future, pp. 9–14. HFES (2016). https://doi.org/10.1177/2327857916051000
10. Li, Y., Wang, X., Burns, C.M.: Modeling automation with cognitive work analysis to support human-automation coordination. J. Cogn. Eng. Decis. Mak. **11**, 299–322 (2017). https://doi.org/10.1177/1555343417709669

11. Li, Y., Wang, X., Burns, C.M.: Improved monitoring performance of financial trading algorithms using a graphical display. Proc. Hum. Factors Ergon. Soc. Annu. Meet. **62**(1), 187–191 (2018). https://doi.org/10.1177/1541931218621045
12. Longo, L.: Designing medical interactive systems via assessment of human mental workload. In 2015 IEEE 28th International Symposium on Computer-Based Medical Systems (CBMS), pp. 364–365. IEEE, June 2015
13. Longo, L., Leva, M.C.: Human Mental Workload: Models and Applications, vol. 726. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61061-0
14. Mijović, P., Milovanović, M., Ković, V., Gligorijević, I., Mijović, B., Mačužić, I.: Neuroergonomics method for measuring the influence of mental workload modulation on cognitive state of manual assembly worker. In: Longo, L., Leva, M.C. (eds.) H-WORKLOAD 2017. CCIS, vol. 726, pp. 213–224. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61061-0_14
15. Moustafa, K., Luz, S., Longo, L.: Assessment of mental workload: a comparison of machine learning methods and subjective assessment techniques. In: Longo, L., Leva, M.C. (eds.) H-WORKLOAD 2017. CCIS, vol. 726, pp. 30–50. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61061-0_3
16. Naikar, N., Moylen, A., Pearce, B.: Analysing activity in complex systems with cognitive work analysis: concepts, guidelines and case study for control task analysis. Theor. Issues Ergon. Sci. **7**, 371–394 (2006). https://doi.org/10.1080/14639220500098821
17. Rasmussen, J.: The role of hierarchical knowledge representation in decision making and system management. IEEE Trans. Syst. Man Cybern. **15**, 234–243 (1985). https://doi.org/10.1109/tsmc.1985.6313353
18. Rizzo, L., Longo, L.: Representing and inferring mental workload via defeasible reasoning: a comparison with the NASA task load index and the workload profile. In: AI^3@AI*IA, 1st Workshop on Advances in Argumentation in Artificial Intelligence co-located with XVI International Conference of the Italian Association for Artificial Intelligence (AI*IA 2017), Bari, Italy, 16–17 November 2017 (2017)
19. Rizzo, L., Dondio, P., Delany, S.J., Longo, L.: Modeling mental workload via rule-based expert system: a comparison with NASA-TLX and workload profile. In: Iliadis, L., Maglogiannis, I. (eds.) AIAI 2016. IAICT, vol. 475, pp. 215–229. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-44944-9_19
20. Smith, K.T.: Observations and issues in the application of cognitive workload modelling for decision making in complex time-critical environments. In: Longo, L., Leva, M.C. (eds.) H-WORKLOAD 2017. CCIS, vol. 726, pp. 77–89. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61061-0_5
21. St. Maurice, J., Burns, C.M.: Using comparative cognitive work analysis to identify design priorities in complex socio-technical systems. In: 2015 International Symposium on Human Factors and Ergonomics in Health Care: Improving the Outcomes, vol. 4, no. 1, pp. 118–123 (2015). https://doi.org/10.1177/2327857915041019
22. Vicente, K.J.: Cognitive Work Analysis: Toward Safe, Productive, and Health Computer-Based Work. LEA, Mahwah (1999)
23. Wickens, C.D.: Mental workload: assessment, prediction and consequences. In: Longo, L., Leva, M.C. (eds.) H-WORKLOAD 2017. CCIS, vol. 726, pp. 18–29. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61061-0_2

Assessing Workload in Human-Machine Teams from Psychophysiological Data with Sparse Ground Truth

David Dearing(✉), Aaron Novstrup, and Terrance Goan

Stottler Henke Associates, Inc.,
1107 NE 45th Street, Suite 310, Seattle, WA 98105, USA
{ddearing, anovstrup, goan}@stottlerhenke.com

Abstract. Data-driven approaches to human workload assessment generally attempt to induce models from a collection of available data and a corresponding ground truth comprising self-reported measures of actual workload. However, it is often not feasible to elicit self-assessed workload ratings with great frequency. As part of an ongoing effort to improve the effectiveness of human-machine teams through real-time human workload monitoring, we explore the utility of transfer learning in situations where there is sparse subject-specific ground truth from which to develop accurate predictive models of workload. Our approach induces a workload model from the psychophysiological data collected from subjects operating a remotely piloted aircraft simulation program. Psychophysiological measures were collected from wearable sensors, and workload was self-assessed using the NASA Task Load Index. Our results provide evidence that models learned from psychophysiological data collected from other subjects outperform models trained on a limited amount of data for a given subject.

Keywords: Workload assessment · Transfer learning · Human-machine teams · Psychophysiological sensors · Human-automation interaction · Machine learning

1 Introduction

Effective human workload assessment techniques have long been sought after by researchers in hopes of preventing fatigue, stress, and other negative influences on performance [1–3]. One particular application of such techniques is to diagnose performance successes and failures in human-machine teams to help identify effective training and design interventions. A wide range of research suggests that problems in such teams are greatly exacerbated by the harmful effects that high cognitive demands can have on human operator performance [4–12]. Many approaches have been proposed for assessing human workload, ranging from theory-driven models to data-driven computational models [13]. In the case of computational models, machine learning can be used to train a model directly from measurable factors—producing a model by finding patterns and relationships between an individual's measurable factors and a corresponding performance measure.

L. Longo and M. C. Leva (Eds.): H-WORKLOAD 2018, CCIS 1012, pp. 13–22, 2019.
https://doi.org/10.1007/978-3-030-14273-5_2

In this paper, we report on our initial efforts to induce such a data-driven workload model in situations where there is only a sparse amount of data for a given human operator. Our research focuses on improving the effectiveness of human-machine teams through real-time human workload monitoring captured by wearable sensors. More specifically, we describe our progress in the context of an ongoing effort to develop an extensible modeling framework and software system for real-time human state assessment in human-machine teaming environments.

The remainder of this paper is organized as follows. Section 2 describes related work in human workload assessment and the challenges of machine learning with sparse data. Section 3 describes the design of a comparative study to evaluate the utility of the transfer learning technique with psychophysiological data for human workload monitoring. Section 4 presents the experimental results. Finally, Sect. 5 presents the conclusions and future work.

2 Related Work

Traditional approaches to producing an index of workload have typically been theory-driven, following a top-down approach that begins with a hypothesis based on existing knowledge and then moves towards the measurement and quantification of the factors believed to influence workload [14–16]. Recently, however, there has been an increased focus on data-driven approaches [13, 17–19]. Unlike theory-driven approaches, these data-driven approaches can induce a workload model bottom-up from data acquired through subjective self-report measures and other measurable factors. In particular, there has been a rise in the use of psychophysiological data to generate these data-driven models, in part because such measures can be captured unobtrusively with wearable sensors and, thus, fully integrated into real-world work environments [20–23]. However, these data-driven approaches still rely on self-assessed workload ratings to serve as the labeled ground truth for training a machine learning classifier. Because these self-reports often require the full cognitive attention of the user, it is often infeasible to elicit these self-assessed ratings with great frequency (i.e., while performing attention-demanding tasks like driving or flying). In such situations, where there is sparse subject-specific ground truth data from which to develop accurate predictive models of workload, a more effective alternative might be to utilize models produced from the labeled data collected from *other* subjects. This technique, called *transfer learning*, has been used in other applications where it is expensive or impossible to collect the needed training data and rebuild the models [24]. Although the psychophysiological data from other subjects may not be completely consistent with a new subject's profile, it may still contain useful information, as people may exhibit similar responses to the same task [25].

3 Design and Methodology

This section describes the design of a comparative study to evaluate different approaches to inducing a workload model from psychophysiological data. In particular, we are interested in the value of transfer learning in situations where there is sparse ground truth for a given human operator (i.e., labeled psychophysiological data) from which to develop accurate predictive models of workload. This would be the case when, for example, a new operator joins a human-machine team and no physiological measures or self-assessed workload ratings have yet to be collected. Our research centers on examining the following hypotheses:

Hypothesis 1: Given a sparse amount of psychophysiological data for a particular human operator, a workload model induced from that sparse data and the data collected from all *other* human operators will outperform a workload model trained using the data for that operator alone.

Hypothesis 2: Given a sufficient volume of psychophysiological data for a particular human operator, an operator-specific workload model will outperform a model trained using data from the all *other* human operators.

3.1 Dataset

We utilized an existing dataset consisting of psychophysiological measures collected as part of a formal study conducted by the Air Force Research Laboratory (AFRL) Human Universal Measurement and Assessment Network (HUMAN) Laboratory. In this study, a total of 13 participants were instrumented with wearable sensors to collect physiological data consisting of respiration measures, electrocardiography (ECG) measures, and electroencephalography (EEG) measures (see Table 1 and [26] for details). Each participant was monitored while operating a remotely piloted aircraft simulation program and workload was self-assessed using the NASA Task Load Index (NASA-TLX) [27].

Table 1. Summary table of dataset features.

Independent Features			
Respiration Features	**ECG Features**	**EEG Features**	
Amplitude	Heart Rate	Electrodes:	F7, F8, T3, T4, Fz, O2
Cycle Time	Filtered Heart Rate	Frequencies:	Alpha, Beta, Delta,
Inspiration Time	Heart Rate Variability		Gamma1, Gamma2,
Duty Cycle	Interbeat Interval		Gamma3
	Heart Rate Trend	Measures:	Power, Saccade

Dependent Feature						
NASA-TLX	Sessions	Trials	Min	Max	High Workload	Low Workload
Composite Score	104	205	2.2	94.4	105 Trials	100 Trials

Over the course of ten days, each participant experienced two training sessions and eight data collection sessions. Each session included two seven-minute trials of a target tracking task that played out along distinct scripted timelines during which the participant manually tracked either one or two targets (among other independent variables to vary the task demand). This resulted in a total of 104 data collection sessions with 205 individual trials (data was either missing or erroneous for three of the trials). When each trial of the tracking task ended, participants were asked to fill out a new NASA-TLX questionnaire. The NASA-TLX scale, which is built upon six factors and their individual weights, has been widely used by human factors researchers over the last four decades. Each 7-min trial of an operator's session is labeled with a single, static NASA-TLX composite measure of perceived workload. The physiological measures were collected constantly throughout each trial.

3.2 Workload Model Training

As part of an ongoing effort to improve the effectiveness of human-machine teams through real-time human workload monitoring, Stottler Henke is developing an extensible modeling framework and software system for real-time human state assessment in human-machine teaming environments. The system, called *Illuminate*, employs machine learning techniques to induce a workload model from psychophysiological data and can—in real-time—update its internal model as new labeled data is made available. This enables us to evaluate *adaptive* models that incrementally incorporate operator-specific data (e.g., after each session; see Fig. 1), thereby addressing complications typically associated with the analysis of biometric data, including: the non-stationarity of psychophysiological data [28] and the initial sparsity (or lack) of operator-specific data. Workload models are trained to predict either high or low workload for an operator, based on the self-reported composite NASA-TLX measure that corresponds with the physiological measures for each trial.

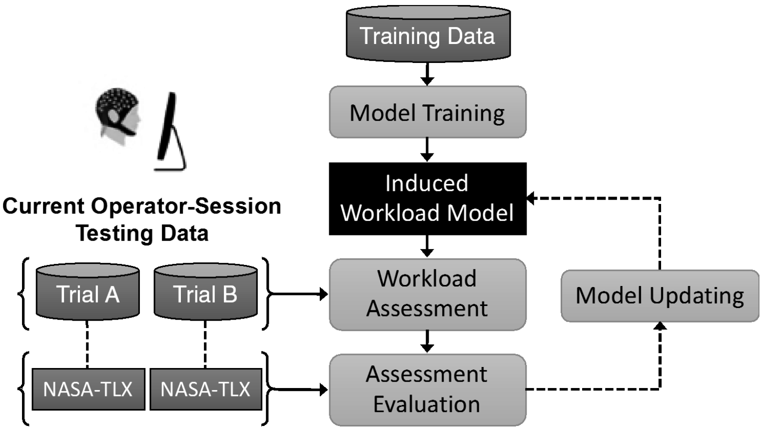


Fig. 1. Diagram of the general experimental workflow for training, evaluating, and updating the workload model for each individual operator session (consisting of two trials).

Data Preparation. Prior to model training, we first normalize the measures on a per-operator basis and designate each trial as either a high or low workload based on the normalized composite NASA-TLX measure (low workload being less than the 50th percentile). Psychophysiological measures are aligned by their individual timestamps and down-sampled to a frequency of 2 Hz with a five-second rolling average, so as to synchronize the measurements collected by the various sensors.

Model Training and Evaluation. Our system uses machine learning classification techniques to induce a workload model for each individual operator and uses that model to assess an operator's workload at a given moment in time. For these experiments, we utilize the Weka implementation of a multinomial logistic regression model with a ridge estimator [29, 30]. For the purposes of our comparative evaluation, we trained models for each operator using three configurations (see Fig. 2):

- *Transfer learning model*: We use a leave-one-out approach in which one operator at a time is taken to be the "current" operator and the data for all of the *other* operators provides the basis for training a transfer learning model.
- *Session-specific model*: For each "current" operator, a session-specific model is trained using the data for all of that operator's *previous* sessions. That is, when evaluating data for the n^{th} session, the corresponding model has been trained on all data for the current operator's first $n - 1$ sessions.
- *Combination model*: As with the session-specific model, the combination model for each "current" operator is updated between each session so that it has been trained on data for all *previous* sessions. It uses model stacking to combine the session-specific model with that operator's corresponding transfer learning model by including the output of the transfer learning model as an additional input when training the session-specific model.

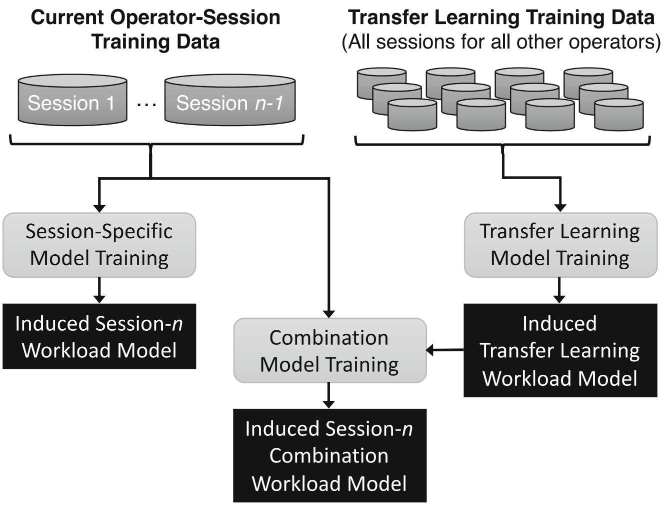


Fig. 2. Diagram outlining the three model training configurations.

These configurations are meant to simulate a human-machine team scenario in which the system initially has a sparse amount of human operator-specific data from which to develop predictive models of workload (i.e., for a new human operator). As the operator completes additional sessions and provides workload feedback (e.g., NASA-TLX measure of perceived workload), the system updates its internal model for subsequent workload assessment. Note that each model is evaluated only on data for sessions 2 – 8—there is no session-specific model for the first session.

4 Results and Discussion

Here we describe the results of our comparative evaluation of the three configurations for model training and evaluation that are described in Sect. 3.2. Because the underlying models are binomial logistic regression models with probabilistic output (as opposed to binary output), when evaluating the models on a given trial we average the workload assessment scores over all instances for that trial. To measure the performance of each model, we calculate the sensitivity and specificity metrics. Sensitivity provides an estimate of how good the model is at predicting a high level of workload, whereas specificity estimates how good the model is at predicting a low level of workload. We plot these measures on a Receiver Operating Characteristic (ROC) curve, a graphical representation wherein the points of the curve are obtained by moving the classification threshold from favoring a correct assessment of low workload to favoring a correct assessment of high workload. Each chart also indicates the point with the "optimal" threshold, maximizing Youden's J statistic [31].

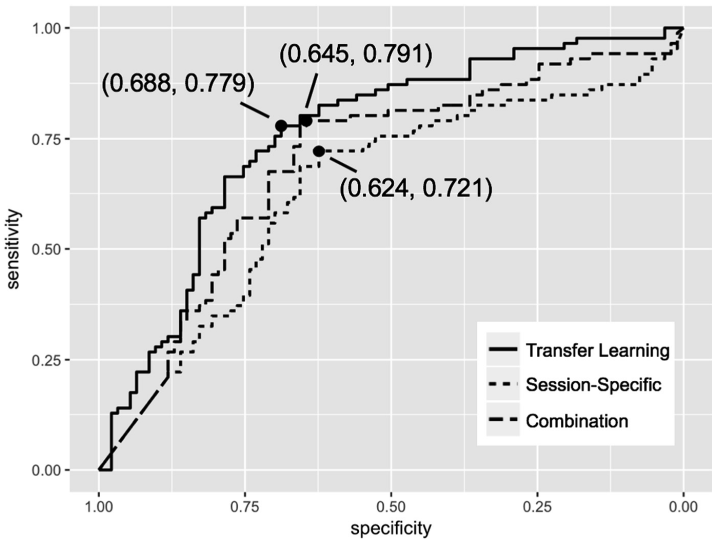


Fig. 3. ROC curve comparing each model configurations, averaged across all 13 operators.

For each of the three evaluation configurations, the chart in Fig. 3 shows the ROC curve averaged across all operators and sessions. Figure 3 shows that the transfer learning model outperforms not only the session-specific model, but also the combination model. These difference between the transfer learning model and the session-specific model is statistically significant (p-value of 0.006), as is the difference between the combination and session-specific models (p-value of 0.005). This provides evidence toward confirming our first hypothesis that psychophysiological data collected from other operators can be used to train a predictive workload model. However, the fact that the transfer learning model outperformed the combination model that utilizes psychophysiological data for the current subject with the results of the transfer learning model was unexpected but not significant (p-value of 0.129).

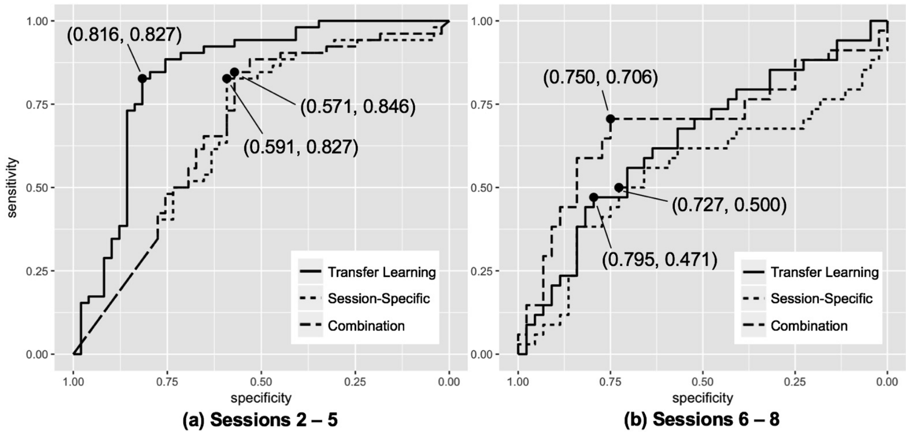


Fig. 4. ROC curves comparing each of the three model configurations, averaged across all 13 operators for (a) the first four sessions and (b) the final three sessions.

To better understand how the session-specific and combinations models performed over time (e.g., with additional training data from later sessions), we partitioned the evaluation data to compare the results of the first four sessions (sessions 2–5) with the results of the final three sessions (sessions 6–8), which resulted in partitions with roughly balanced class labels. As the ROC curves in Fig. 4 illustrate, the combination model actually outperforms the transfer learning model for sessions 6–8 (i.e., when there are four or more sessions of labeled training data available for each operator). Comparing the optimal points on each curve, we can see that the transfer learning model's performance drops during the later sessions (from a J of 0.643 to 0.266), whereas the performance of the combination model improves in the later sessions (from a J of 0.418 to 0.456).

5 Conclusion and Future Work

In this paper, we have described a comparative evaluation to test our hypothesis that transfer learning is useful in situations where there is insufficient subject-specific data to develop accurate predictive models of human workload. Our results provide evidence that models learned from psychophysiological data collected from other subjects outperform models trained on a limited amount of data for a given subject (*Hypothesis 1*). More specifically, with little or no data, a transfer learning model trained on all other-subject data performed best. Once a sparse amount of subject-specific data was available, a combination model induced from the output of the transfer learning model and the subject-specific psychophysiological measures generally outperformed the transfer learning model alone.

There remain several items to be answered by future work as well as by our own ongoing research. First, this evaluation does not answer the question as to at what point a session-specific model (i.e., induced only from a subject's own data) outperforms the transfer learning model (*Hypothesis 2*). Future work is needed to inspect which features contribute most to the performance of each model and how those features change across the three configurations. We are also left questioning why the transfer learning model performs worse for the later sessions. Does the transfer learning model have trouble due to the non-stationarity of psychophysiological data (whereas the other models adapt as new subject-specific data is collected)? Or, alternatively, is this an artifact of this particular dataset? Additional studies that collect data over more trials would be necessary to answer these questions.

Another topic for future research would be to vary the dividing line between what constitutes a high and low workload. Our evaluation uses the median of the normalized composite NASA-TLX measure as a simple and straightforward dividing line between high and low workload. Depending on the task and application domain, it may be more appropriate to raise (or lower) that threshold. Alternatively, to account for potential learning effects across sessions, a per-subject adaptive threshold may yield better results. The results of our comparative evaluation highlight the threshold for each model that produces the best balance of sensitivity and specificity. However, depending on the target application, it may be more appropriate to favor specificity or sensitivity so as to more accurately predict a high or low workload, respectively.

Lastly, a more accurate model might be produced by first identifying a relevant subset of other subjects within the transfer learning data. In particular, if individual differences are high (as is often the case with psychophysiological data), a more accurate model might be induced based only on data collected from people who appear to have similar physiological responses. This is something we plan to explore in our ongoing work to improve the effectiveness of human-machine teams through real-time human workload monitoring.

Acknowledgments. This material is based upon work supported by the United States Air Force under Contract No. FA8650-15-C-6669. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the United States Air Force.

References

1. Hancock, P.A.: Whither workload? Mapping a path for its future development. In: Longo, L., Leva, M.C. (eds.) H-WORKLOAD 2017. CCIS, vol. 726, pp. 3–17. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61061-0_1
2. Longo, L.: Mental workload in medicine: foundations, applications, open problems, challenges and future perspectives. In: 2016 IEEE 29th International Symposium on Computer-Based Medical Systems (CBMS), pp. 106–111 (2016). https://doi.org/10.1109/cbms.2016.36
3. Wickens, C.D.: Mental workload: assessment, prediction and consequences. In: Longo, L., Leva, M.C. (eds.) H-WORKLOAD 2017. CCIS, vol. 726, pp. 18–29. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61061-0_2
4. Balfe, N., Crowley, K., Smith, B., Longo, L.: Estimation of train driver workload: extracting taskload measures from on-train-data-recorders. In: Longo, L., Leva, M.C. (eds.) H-WORKLOAD 2017. CCIS, vol. 726, pp. 106–119. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61061-0_7
5. Chen, J.Y., Haas, E.C., Barnes, M.J.: Human performance issues and user interface design for teleoperated robots. IEEE Trans. Syst. Man Cybern. Part C (Appl. Rev.) **37**(6), 1231–1245 (2007). https://doi.org/10.1109/tsmcc.2007.905819
6. Edwards, T., Martin, L., Bienert, N., Mercer, J.: The relationship between workload and performance in air traffic control: exploring the influence of levels of automation and variation in task demand. In: Longo, L., Leva, M.C. (eds.) H-WORKLOAD 2017. CCIS, vol. 726, pp. 120–139. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61061-0_8
7. Fan, J., Smith, A.P.: The impact of workload and fatigue on performance. In: Longo, L., Leva, M.C. (eds.) H-WORKLOAD 2017. CCIS, vol. 726, pp. 90–105. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61061-0_6
8. Hancock, P.A., Williams, G., Manning, C.M.: Influence of task demand characteristics on workload and performance. Int. J. Aviat. Psychol. **5**(1), 63–86 (1995). https://doi.org/10.1207/s15327108ijap0501_5
9. Longo, L.: Experienced mental workload, perception of usability, their interaction and impact on task performance. PLoS ONE **13**(8), e0199661 (2018). https://doi.org/10.1371/journal.pone.0199661
10. Orasanu, J.M.: Shared problem models and flight crew performance. In: Johnston, N., McDonald, N., Fuller, R. (eds.) Aviation Psychology in Practice. Ashgate Publishing Group, Aldershot (1994). https://doi.org/10.4324/9781351218825
11. Smith, A.P., Smith, H.N.: Workload, fatigue and performance in the rail industry. In: Longo, L., Leva, M.C. (eds.) H-WORKLOAD 2017. CCIS, vol. 726, pp. 251–263. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61061-0_17
12. Tong, S., Helman, S., Balfe, N., Fowler, C., Delmonte, E., Hutchins, R.: Workload differences between on-road and off-road manoeuvres for motorcyclists. In: Longo, L., Leva, M.C. (eds.) H-WORKLOAD 2017. CCIS, vol. 726, pp. 239–250. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61061-0_16
13. Moustafa, K., Luz, S., Longo, L.: Assessment of mental workload: a comparison of machine learning methods and subjective assessment techniques. In: Longo, L., Leva, M.C. (eds.) H-WORKLOAD 2017. CCIS, vol. 726, pp. 30–50. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61061-0_3
14. Kantowitz, B.H.: Mental workload. In: Hancock, P.A. (ed.) Advances in Psychology, vol. 47, pp. 81–121. North-Holland (1987). https://doi.org/10.1016/s0166-4115(08)62307-9

15. Longo, L., Barrett, S.: A computational analysis of cognitive effort. In: Nguyen, N.T., Le, M. T., Świątek, J. (eds.) ACIIDS 2010. LNCS (LNAI), vol. 5991, pp. 65–74. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-12101-2_8
16. Moray, N. (ed.): Mental Workload: Its Theory and Measurement, vol. 8. Springer, New York (2013). https://doi.org/10.1007/978-1-4757-0884-4
17. Appriou, A., Cichocki, A., Lotte, F.: Towards robust neuroadaptive HCI: exploring modern machine learning methods to estimate mental workload from EEG signals. In: Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems, p. LBW615. ACM (2018). https://doi.org/10.1145/3170427.3188617
18. Wilson, G.F., Russell, C.A.: Real-time assessment of mental workload using psychophysiological measures and artificial neural networks. Hum. Factors **45**(4), 635–644 (2003). https://doi.org/10.1518/hfes.45.4.635.27088
19. Zhang, J., Yin, Z., Wang, R.: Recognition of mental workload levels under complex human–machine collaboration by using physiological features and adaptive support vector machines. IEEE Trans. Hum.-Mach. Syst. **45**(2), 200–214 (2015). https://doi.org/10.1109/thms.2014.2366914
20. Prinzel III, L.J., Parasuraman, R., Freeman, F.G., Scerbo, M.W., Mikulka, P.J., Pope, A.T.: Three experiments examining the use of electroencephalogram, event-related potentials, and heart-rate variability for real-time human-centered adaptive automation design. Report TP-2003-212442, NASA, Langley Research Center, Hampton (2003)
21. Roscoe, A.H.: Assessing pilot workload. Why measure heart rate, HRV and respiration? Biol. Psychol. **34**(2–3), 259–287 (1992). https://doi.org/10.1016/0301-0511(92)90018-p
22. Veltman, J.A., Gaillard, A.W.K.: Physiological workload reactions to increasing levels of task difficulty. Ergonomics **41**(5), 656–669 (1998). https://doi.org/10.1080/001401398186829
23. Verwey, W.B., Veltman, H.A.: Detecting short periods of elevated workload: a comparison of nine workload assessment techniques. J. Exp. Psychol.: Appl. **2**(3), 270 (1996). https://doi.org/10.1037/1076-898X.2.3.270
24. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Trans. Knowl. Data Eng. **22**(10), 1345–1359 (2010). https://doi.org/10.1109/tkde.2009.191
25. Wu, D., Lance, B.J., Parsons, T.D.: Collaborative filtering for brain-computer interaction using transfer learning and active class selection. PLoS ONE **8**(2), e56624 (2013). https://doi.org/10.1371/journal.pone.0056624
26. Hoepf, M., Middendorf, M., Epling, S., Galster, S.: Physiological indicators of workload in a remotely piloted aircraft simulation. In: 18th International Symposium on Aviation Psychology, pp. 428–433. Curran, Dayton (2015)
27. Hart, S.G., Staveland, L.E.: Development of NASA-TLX (task load index): results of empirical and theoretical research. In: Advances in Psychology, vol. 52, pp. 139–183. North-Holland (1988). https://doi.org/10.1016/s0166-4115(08)62386-9
28. Christensen, J.C., Estepp, J.R., Wilson, G.F., Russell, C.A.: The effects of day-to-day variability of physiological data on operator functional state classification. NeuroImage **59**(1), 57–63 (2012). https://doi.org/10.1016/j.neuroimage.2011.07.091
29. Le Cessie, S., Van Houwelingen, J.C.: Ridge estimators in logistic regression. Appl. Stat. **41**(1), 191–201 (1992). https://doi.org/10.2307/2347628
30. Witten, I.H., Frank, E., Hall, M.A., Pal, C.J.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, Burlington (2011). https://doi.org/10.1016/c2009-0-19715-5
31. Youden, W.J.: Index for rating diagnostic tests. Cancer, **3**(1), 32–35 (1950). https://doi.org/10.1002/1097-0142(1950)3:1<32::aid-cncr2820030106>3.0.co;2-3

The Evolution of Cognitive Load Theory and the Measurement of Its Intrinsic, Extraneous and Germane Loads: A Review

Giuliano Orru and Luca Longo(✉)

School of Computing, College of Health and Sciences,
Dublin Institute of Technology, Dublin, Ireland
Luca.Longo@dit.ie

Abstract. Cognitive Load Theory has been conceived for supporting instructional design through the use of the construct of cognitive load. This is believed to be built upon three types of load: intrinsic, extraneous and germane. Although Cognitive Load Theory and its assumptions are clear and well-known, its three types of load have been going through a continuous investigation and re-definition. Additionally, it is still not clear whether these are independent and can be added to each other towards an overall measure of load. The purpose of this research is to inform the reader about the theoretical evolution of Cognitive Load Theory as well as the measurement techniques and measures emerged for its cognitive load types. It also synthesises the main critiques of scholars and the scientific value of the theory from a rationalist and structuralist perspective.

Keywords: Cognitive Load Theory, Cognitive load types · Intrinsic load, Extraneous load, Germane load · Measures · Instructional design · Efficiency

1 Introduction

The construct of Cognitive Load (CL) is strictly related to the construct of Mental Workload (MWL). The former has evolved within Educational Psychology [1], while the latter within Ergonomics and Human Factors [2]. Despite their independent evolution within different disciplines, both are based upon the same core assumption: the limitations of the human mental architecture and the cognitive capacities of the human brain and its working memory [3, 4]. In a nutshell, as professor Wickens suggested [5], mental workload is equivalent to the amount of mental resources simultaneously elicited by a human during the execution of a task. In order to achieve an optimal performance, the working memory limits should not be reached [3, 4]. If this occurs, the mental resources are no longer adequate to optimally execute the underlying task. Within Ergonomics, the construct of Mental Workload has evolved both theoretically and practically. A plethora of ad-hoc definitions exist as well as several domain-dependent measurement techniques, measures and applications [1]. While abundance of research exists, the science of Mental Workload is still in its infancy because any of the proposed measures can generalise the construct itself. Similarly, within Educational

L. Longo and M. C. Leva (Eds.): H-WORKLOAD 2018, CCIS 1012, pp. 23–48, 2019.
https://doi.org/10.1007/978-3-030-14273-5_3

Psychology, despite Cognitive Load Theory (CLT) is one of the most invoked learning theory for supporting instructional design [6], research on how to develop highly generalisable measures of Cognitive Load is limited. Also, it is unclear how its three types of load – intrinsic, extraneous and germane – can be measured and how they interact with each other. The aim of this paper is to provide readers with the theoretical elements underpinning the construct of Cognitive Load. This is done from an evolutionary perspective of the measurement techniques and measures emerged from the three types of load, accompanied with a critical discussion of their scientific value.

The remainder of the paper is structured as follows. Section 2 presents the key theoretical elements and assumptions of Cognitive Load Theory, as appeared in the literature. Section 3 focuses on a review of the measurement techniques and measures emerged for its intrinsic, extraneous and germane loads. Section 4 builds on this review by emphasising the open debate on the scientific value of CLT. Section 5 highlights new perspectives and research on CLT and the reconceptualization of its cognitive load types, as recently emerged in the literature. Section 6 summarises this study with final remarks suggesting novel research directions.

2 Cognitive Load Theory

Cognitive Load Theory (CLT) is a cognitivist learning theory aimed at supporting instructors in the development of novel instructional designs aligned with the limitations of the human cognitive architecture. In a nutshell, this architecture is the human cognitive system aimed at storing information, retrieving and processing it for reasoning and decision making [7]. CLT is based upon the assumption of *active processing* that views the learner as actively engaged in the construction of knowledge [8]. In other words, learners are actively engaged in a process of attention to relevant material and its organisation into coherent structures that are integrated with prior knowledge [9]. Another premise of CLT is the *dual-channel* assumption by which processing of information occurs in two distinct channels: an auditory and a verbal channel. The former processes auditory sensory input and verbal information while the latter processes visual sensory inputs and pictorial representations [10]. An essential component of this architecture is its memory that can store information for short and long term. According to another premise, *the limited capacity* assumption of CLT, the former memory, also referred to working memory, is conscious and limited, while the latter is unconscious and unlimited [2]. Baddeley [3] and Paivio [4], following Miller proposal [7], support the view that when working memory has to deal with new information, it can hold just seven chunks at a time. However, if these chunks are related and if they have to be processed, human beings are capable to handle just two or three at the same time [11]. Expanding the capacity of working memory coincides with learning [2]. Learning take places by transferring pieces of information from working memory to long term memory [3, 4]. According to Schema Theory, this transfer of information allows the construction of knowledge, in long term memory, in the form of schema [12]. To construct a schema means to relate different chunks of information from a lower level to a higher level of complexity and to hold them as a single unit that can be understood as a single chunk of information [12]. In turn, schema can be retrieved to solve a problem, a task, or more generally to answer a question in educational contexts. Schema construction is believed to reduce the load in working memory [2]. The expansion of long term memory can be achieved by a reduction of the load of working memory. Leaving sufficient cognitive resources in working memory to process new information is one of the core objectives of educational instructional design. In fact, if the amount of information that has to be held in working memory lies within its limits, the learning phase is facilitated. Contrarily, if the amount of information overcomes these limits, an overload situation occurs and the learning phase is hampered (Fig. 1).

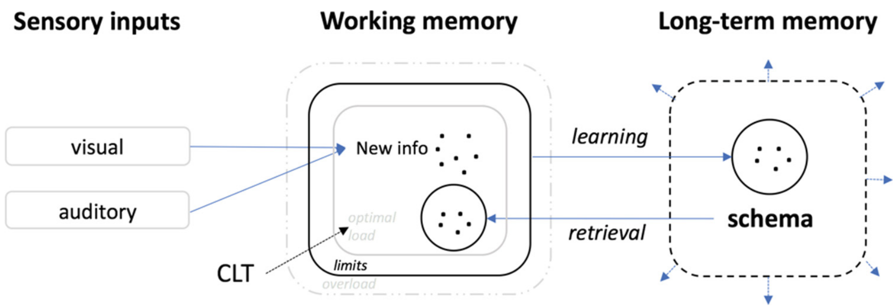


Fig. 1. A representation of the mental architecture and the role of Cognitive Load Theory (CLT) in connection to working memory and schema construction

A core construct within CLT is Cognitive Load (CL), believed to be multidimensional. Intuitively it can be defined as the mental cost imposed by an underlying cognitive task on the human cognitive system [13]. It is possible to distinguish two types of factors that can interact with cognitive load: causal and assessment factors (Fig. 2). The formers affect cognitive load while the latter are affected by cognitive load. The causal factors include:

- features of the task (T) such as structure, novelty and pressure;
- the features of the environment (E) such as noise and temperature where a task (T) is executed and their interaction (ExT);
- the characteristic of a learner (L) such as capabilities, cognitive style and prior knowledge;
- the interaction between environment and learner characteristics (ExL);
- the interaction between task, environment, learner's characteristics Ex(TxL).

The assessment factors can be conceptualised with three dimensions: mental load, mental effort and mental performance. Mental load is imposed by the task (T) and/or by demands from the environment. It is a task-centred dimension, independent of the subject, and it is considered constant. Mental effort is a human-centred dimension that reflects the amount of controlled processing (capacity or resources allocated for task demands) in which the individual is engaged with [13]. It is affected by the task-environment interaction (ExT), the subject characteristics interaction with the environment (ExL) and the interaction of the learner with the task in the environment (Ex(TxL)). Similarly, the level of mental performance is affected by the factors that affect mental effort [4]. Other factors might affect cognitive load [14, 15] and research in the field has not produced a comprehensive list yet [16].

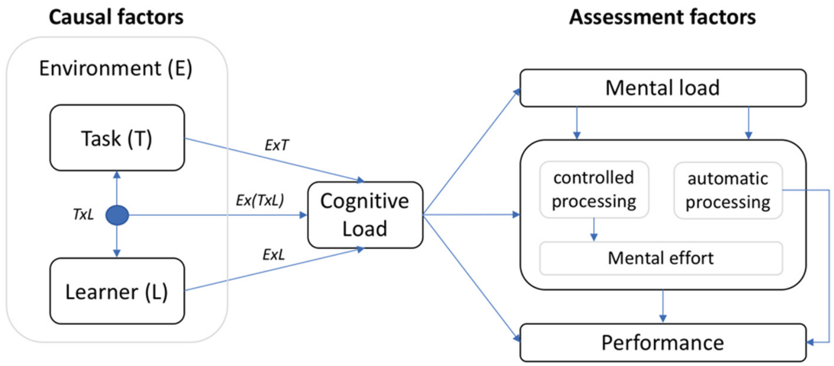


Fig. 2. Causal factors and assessment factors according to [41].

Starting from the research of Halford et al. [17] on the difficulty in processing information with multiple elements at the same time during problem solving, Sweller defined the degree of complexity of these elements as 'element interactivity' [18]. Starting from this definition, two types of cognitive load has emerged: the *intrinsic* and the *extraneous* loads. Intrinsic load refers to the numbers of elements that must be processed simultaneously in working memory (element interactivity) for schema construction. 'This type of load cannot be modified by instructional interventions because it is intrinsic to the material being dealt with. Instead, extraneous cognitive load is the unnecessary cognitive load and can be altered by instructional interventions' [2]. Sweller stated that the basic goal of Cognitive Load Theory is the reduction of extraneous load: this is a type of ineffective load that depends on the instructional techniques provided by the instructional format to complete a task [2]. This view is supported by Paas and colleagues that refers to extraneous load as the cognitive effect of instructional designs that hamper the construction of schema in working memory [19]. Beside, intrinsic and extraneous, Sweller defined another type of load: the *germane* load [2]. This is the extra effort required for learning (schema construction). It is possible to use this effort when intrinsic and extraneous loads leave sufficient working memory resources. This extra effort increases cognitive load, but it is connected to learning, thus, it facilitates schema construction. Germane load is the effective cognitive load and it is the result of those beneficial cognitive processes such as abstractions and elaboration that are promoted by 'good' instructional designs [20]. Reducing extraneous load and improving germane load by developing schema construction and automation should be the main goal of the discipline of instructional design. The three types of load emerged within Cognitive Load Theory, and their role, can be summarised in Fig. 3.

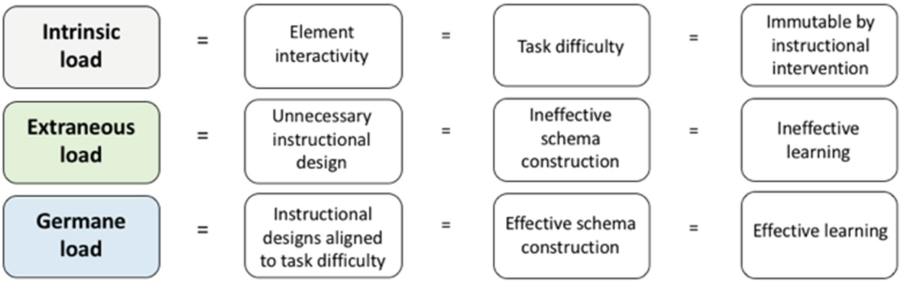


Fig. 3. Definitions and role of the cognitive load types of Cognitive Load Theory.

Sweller and colleagues, with their attempt to define cognitive load within the discipline of Educational Psychology and for instructional design, believed that the three types of load are additive. This meant that the total cognitive load experienced by a learner in working memory while executing a task, is the sum of the three types of load, these being independent sources of load [2] (Fig. 4).

Fig. 4. Additive definition of overall cognitive load.

Figure 5 depicts the relationship between the three types of cognitive load, as proposed in [21]. In condition A (overload), cognitive load exceeds the limits of the working memory of the learner due to an increment in the extraneous load. In turn errors are more frequent, longer task execution times occur, sometimes even leading to the inability to perform an underlying task. In condition B there is spare working memory capacity and the learners can perform optimally on an underlying task. With spare capacity, CLT proposes to increase the germane load in order to activate learning tasks, as in condition C.

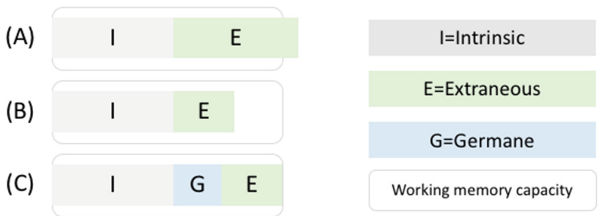


Fig. 5. Relationship between the three types of cognitive load

3 Cognitive Load Types, Measurement Techniques and Measures

The definition of the types of load within Cognitive Load Theory (CLT) are supported by empirical studies that have used three different classes of measures:

- *task performance measures* such as error rates, learning times and secondary task measures;
- *subjective measures* such as self-reporting questionnaires and rating scales;
- *physiological measures* such as eye movements and physical body responses.

Within Educational Psychology, the focus has always been on the first two classes: task performance and subjective measures. The reason is intuitive since physiological measurement techniques require special equipment to gather data, trainer operators to use this equipment and they are intrusive, most of the time not suitable for empirical experiments in typical classrooms. Additionally, evidence suggests they did not prove sufficient sensitivity to differentiate the three cognitive load types [13] envisioned in CLT. As a consequence, the next sections mainly focus on research studies that employed task performance and subjective measures.

Miwa et al. [21] developed a task-performance based method for cognitive load measurement built upon the *mental chronometry paradigm* [22], in line with the tri-archic view of load [2]. The mental chronometry assumes that reaction time can reveal the quantity of intrinsic, extraneous and germane loads coming from the corresponding cognitive processes. The hypothesis behind their experiment is that if, in the 8×8 Reversi game (Fig. 4), the three types of cognitive load are manipulated by changing the presentation of the information to the players, it is possible to measure them by observing their reaction time of players between movements of discs on the board. To manipulate the intrinsic load, on one hand, an advisor computer agent provides some hints to participants on the possible subsequent move (low intrinsic load condition, Fig. 6 left). On the other hand, in another condition, no hints are provided (high intrinsic load condition, Fig. 6 right). In addition to this, in order to manipulate extraneous load, the white and black discs are changed with two different letters from the Japanese alphabet. Since these 2 letters are perceptually similar, they are expected to lead to higher perception and understanding exerted by participants. Eventually, the germane load is manipulated by altering the instructions presented to participants. In order to exert more germane load, each participant is requested to report, after the game, the heuristics learnt to play it. According to this paradigm, if the reaction time, in high intrinsic load conditions, is longer than the reaction time, in low intrinsic load conditions, then it can be considered as a valid indicator of intrinsic load. Here, learning corresponds to the development of effective strategies for moving discs that can lead a participant to win the game. These strategies imply that players control and regulate their cognitive processing by meta cognitive perspectives, thus increasing their germane load [21].

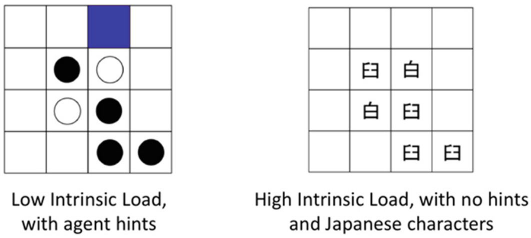


Fig. 6. Low and high extraneous load conditions in the 8 × 8 Reversi game [35].

The research attempt by Miwa and colleagues [21] is indeed useful to investigate the discrimination between the three types of load and to provide guidelines on how to design experiments that contribute to the definition of cognitive load. Their preliminary findings suggest the three types of load are separable. However, it was executed in a highly controlled environment and not in more natural settings such as in a typical classroom, thus limiting the generalisability of their findings.

3.1 Subjective Measures of Cognitive Load

Gerjets et al. [23] proposed two experiments on the use of hypermedia environments for learning on the topic of probability theory. In the first experiment, the validity of multimedia principles [10] in hypermedia environments has been tested. In the second experiment, an analysis of the ability of learners to impact their performance according to their prior experience was performed. A subjective 9-point Likert scale is employed to measure the three types of load during learning (Table 1).

Table 1. Subjective rating scale to measure cognitive load types from [23]. Each item has to be rated on a 9 point Likert scale (1 = extremely easy, 9 = extremely difficult)

Type of load	Scale
Intrinsic load	How easy or difficult do you consider probability theory at this moment?
Extraneous load	• How easy or difficult is it for you to work with the learning environments? • How easy or difficult is it for you to distinguish important and unimportant information in the learning environments? • How easy or difficult is it for you to collect all the information that you need in the learning environment?
Germane load	Indicate the amount of effort you exerted to follow the last example

Depending on the prior knowledge of learners, high intrinsic and extraneous loads should lead to poor learning outcomes while high germane load should lead to good learning outcomes. Unfortunately, in this study no evidence of this connection has been found. Consequently, authors claim that subjective rating scales are valid if learners are able to distinguish the types of cognitive load. In other words, in order to be sensitive to the differences in loads, learners should be aware of the cognitive process connected to the experienced load. To achieve this, training learners on Cognitive Load Theory can facilitate their understanding of the three types of load. However, this is not an easy condition to achieve. In fact, for instance, the level of difficulty (intrinsic load) could be due to the poor instructional design that increase the extraneous load, or due to the natural complexity of an underlying learning task. A novice learner could find this distinction really hard to understand and could not be able to comprehend if the own difficulty in learning can be attributed to the instructional design (extraneous load) or the complexity of the task (intrinsic load).

Corbalan et al. [24] hypothesises that, in order to prevent cognitive overload, it is possible to adapt task difficulty and the support of each newly selected learning task to the previous knowledge and experience of a learner and his/her perceived task load [24]. This can be done by employing some external agent, the learner him/herself, or both. The hypothesis was tested by employing two subjective rating scales, one for task load and one for germane load (as per Table 2). This hypothesis was tested by performing an empirical 2 $\times$ 2 factorial design experiment with health sciences students. The design variables where the factors adaptation (absence or presence of the agent) and the control over task selection (program control or shared control program/learner).

Table 2. Measurements of task and germane load on a 7-point scale from [24] (1 = extremely low, 7 = extremely high)

Type of load	Scale
Task load	Rate your effort to perform the task
Germane load	Evaluate the effort invested in gaining understanding of the relationship dealt with in the simulator task

Findings suggest that, on one hand, the presence of adaptation delivered more efficient learning and task involvement. On the other hand, shared control produced a higher task involvement when compared to program task selection. Learning here refers to good learning outcomes and lower effort exerted in the underlying learning tasks. Task involvement refers to good learning outcomes and higher effort exerted in the learning task. Both the cases prevented cognitive overload [24].

Ayres hypothesised that, by maintaining the extraneous and the germane loads constant, students can identify changes in the 'element interactivity' within problems by means of subjective measures, and thus successfully quantify the intrinsic cognitive load [25]. In his study, extraneous and germane loads are maintained constant by not providing any instructional hint. Learners had to solve a set of four brackets-expansion problems without any explicit instructions (source of extraneous load) and without any didactic feedback (source of germane load). The bracket-expansion problems required a series of four operations in which the level of difficulty increased. Under this instructional condition, any change in the overall cognitive load is due to change in the element interactivity (source of intrinsic load). Intrinsic load is measured by a subjective measure, as depicted in Table 3. After each operation, learners had to rate its difficulty. The hypothesis is that higher intrinsic load should correspond to more errors.

Table 3. Subjective rating scale of intrinsic load on a 7-point scale from [1] (1 = extremely easy, 7 = extremely difficult)

Type of load	Scale
Intrinsic load	How easy or difficult you found each calculation?

The authors tested their hypothesis with two experiments. In the first, students had low prior mathematical knowledge while in the second, participants had a wider range of mathematical skills. In the first experiment, students could recognise task difficulty since subjective intrinsic load was highly correlated with the errors committed by themselves. In the second experiment, although students did not commit many mistakes, they still could detect differences in task difficulty. These findings support the high sensitivity and reliability of the employed self-reporting measure. Additionally, the takeaway of this study is that, by keeping constant two sources of load out of three, it is possible to get a measure of the remaining dependent load. Transitively, it turns out that, by keeping constant the extraneous and the intrinsic loads, any change in cognitive load, irrespective of the measurement technique employed, corresponds to variations in the dependent variable, the germane load.

Gerjets et al. [26] investigated how to enhance learning through a comparison of two instructional designs on the same topic: how to calculate the probability of a complex event. The first design condition included worked out examples, while the latter included modular worked examples. To measure the experienced cognitive load of learners in each condition, a modified version of the NASA-Task Load Index was (Table 4) [27]. Readers are referred to [27] for the original version.

Table 4. Modified version of the NASA-TLX from [26] where each scale ranged from 0 to 100 (low level to high level)

Type of load	Scale
Task demands (intrinsic load)	How much mental and physical activity did you require to accomplish the learning task? (Thinking, deciding, calculating, remembering, looking, searching)
Effort (extraneous load)	How hard did you have to work to understand the contents of the learning environment?
Navigational demand	How much effort did you invest to navigate the learning environment?
Understanding	How successful did you feel in understanding the contents?
Stress	How much stress did you experience during learning?

Within Cognitive Load Theory, Sweller [2] stated that task demand is caused by the degree of element interactivity of the task (intrinsic load), while the effort is exerted to achieve an effective understanding of the instructional material. The navigational demands are related to those activities not strictly directed to learning. In this line, Gerjets and colleagues stated that the scale for task demands is aimed at quantifying the intrinsic load of the instructional material, the effort scale at quantifying the germane load and the scale for navigational demands is aimed at quantifying the extraneous load [26]. The hypothesis of the experiment is that the modular presentation of worked examples can increase the germane load more than their molar – as a whole – presentation. Unfortunately, findings did not provide evidence about any increment of the germane load. As a possible interpretation, the authors suggested that the instructional explanations provided during the task to increase the germane load, and the self-explanation derived using worked out examples, created a redundant information and the illusion of understanding hampered the learning instead of improving it. However, in a prior experiment, Gerjets et al. [26] have successful demonstrated that, in an example based learning, the modular presentation of worked examples can actually reduce the intrinsic load and improves the germane load more than the molar presentation of the same problem. The modular presentation provides a part-whole sequencing of the solution procedures whereas the molar presentation provides the solutions of the procedures as a whole. The segmentation of the presentation of the worked example led to a decrease in the degree of interactivity as well as in the number of simultaneous items. In turn, this led to a decrease in the intrinsic load. According to the authors, these findings are more relevant to novice learners, whereas the same instructional design could be redundant for more expert learners because the degree of their expertise increases. Consequently, in the case of expert learners, the molar presentation of solution procedures is a more appropriate instructional design. The modified version of the NASA-TLX, employed in this study, has been applied also in [28]. Here, authors focused on the effects of different kinds of computer-based graphic representations in connection to the acquisition of problem-solving skills in the context of probability theory. Despite different experiments, [26, 29] and [28] did not provide evidence on the reliability and validity of the subjective rating scale employed. Therefore, it can be only hypothesised that this scale is sensitive to the three types of load conceived within CLT.

Galy et al. [30] tested the additivity between the intrinsic, the extraneous and the germane loads by manipulating three factors believed to have an effect on each of them. In detail, this study assumed that task difficulty is an indicator of intrinsic load, time pressure of extraneous load and the level of alertness of germane load. The effect on the experienced overall cognitive load is connected to the manipulation of the extraneous and intrinsic loads which are respectively estimated by the self-reporting of notions of tension (time pressure) and mental effort (task difficulty). The level of alertness is measured by the French paper-and-pencil version of the Thayers's Activation-Deactivation Checklist [31]. Questions are listed in Table 5. For each word in the deactivation list, each student had to tick one from the "not at all", "don't know", "little" and "much" labels. These labels are respectively mapped to weights (1, 2, 3 and 4). The responses were counted up to have a measure of four factors: general activation (GA), deactivation sleep (DS), high activation (HA), and general deactivation (GD). The GA/DS ratio yielded an alertness index.

Table 5. Self report scales of cognitive loads types from [30]. Intrinsic and extraneous load are in the scale 0 to 10 (low time pressure/mental effort to high effort/considerable effort)

Type of load	Scale
Intrinsic load	Rate the mental effort (task difficulty) you experienced during the task
Extraneous load	Rate the tension (time pressure) you experienced during the task
Germane load	Select one of the following responses ("not at all", "don't know", "little" and "much") for each of 20 listed adjectives: active, energetic, vigorous, full of, lively, still, quiet, placid, calm, at rest, tense, intense, clutched up, fearful, jittery, wide-awake, wakeful, sleepy, drowsy, tired

The experimental task consisted of a memory recalling activity with 2 digit numbers (low difficulty) or 3 digit numbers (high difficulty) in four conditions: low difficulty and low time pressure, low difficulty and high time pressure, high difficulty and low time pressure, high difficulty and high time pressure. The difference in cognitive load due to variations in task difficulty and time pressure with respect to the different levels of alertness can be taken as an indicator of differences in the contribution of germane load. In low difficulty and low time pressure conditions, germane load is believed to be substantially inexistent, but in high difficulty and high time pressure conditions, it is assumed that the learner has to employ specific strategies to execute the memory task and thus generating germane load. Authors believed that germane load, as a function of alertness, corresponds to the subject's capability to select strategies to be employed while performing the learning task. However, the implementation of these strategies is determined by the amount of free cognitive resources determined by task difficulty and time pressure [30]. Consequently, the authors claimed that alertness is a germane load factor depending on the quantity of working memory resources left by the intrinsic and extraneous load experienced.

Leppink et al. [32] developed a new instrument for the measurement of intrinsic, extraneous and germane loads. The authors consider the critique of Kalyunga et al. [33] about the expertise reversal effect and its consequences on the learning and on the different types of load. According to this, the same instructional feature may be associated with germane load for a learner and with extraneous load for another learner, depending on the level of expertise and on the level of prior knowledge. To develop a more sensitive instrument to detect changes in cognitive load types, they proposed a multi-item questionnaire (Table 6). Authors conducted experiments in four lectures of statistics, asking to rate difficult or complex formulas, concepts and definitions using the scales in Table 6. In a number of studies, Leppink and colleagues verified 7 hypotheses regarding the reliability of the new instrument compared with other instruments, used in the past, to measure intrinsic load [25], extraneous load [34], germane load [35] and for overall cognitive load [36]. They also tested five hypotheses connected to the expected relationship between prior knowledge and intrinsic load, and between prior knowledge and learning outcomes. Through an exploratory analysis, it has emerged that the reliability of the rating scale was positive, the extraneous load and the germane load elements were negatively correlated and the elements that were supposed to measure intrinsic load were not correlated to germane load.

Table 6. Multi-subjective rating scales of cognitive load types from [32] in the scale 0 to 10 (0 = not at all, 10 = completely).

Type of load	Scale
Intrinsic load	• The topic/topics covered in the activity was/were very complex • The activity covered formulas that I perceive as very complex • The activity covered concepts and definitions that I perceived as very complex
Extraneous load	• The instruction and/or explanation during the activity were very unclear • The instruction and/or explanation were, in terms of learning, very ineffective • The instruction and/or explanations were full of unclear language
Germane load	• The activity really enhanced my understanding of the topic(s) covered • The activity really enhanced my knowledge and understanding of statistics • The activity really enhanced my understanding of the formulas covered • The activity really enhanced my understanding of concepts, definitions

Eventually, the elements that were expected to measure intrinsic load had moderate correlation with extraneous load. The validity of the scales was verified by comparing the subjective ratings with the learning outcomes assessed by a performance test. As hypothesised, a high prior knowledge corresponded to a low intrinsic load. Extraneous cognitive load was higher when a problem was solved by an unfamiliar format and germane load was higher when a problem was solved by a familiar format. There is partial evidence that higher germane load, as measured by multiple subjective scales, lead to higher results on post-task test performance.

Leppink and colleagues [37] criticised their own previous study [32] mentioning the uncertainty of their multiple subjective rating scales to represent the three different types of cognitive load. The main reasons of their critique are three: (1) the correlation between germane load and the learning outcomes, in the task performance, was lower than expected and not statistically relevant (2) the previous experiments were all focused on a single topic, namely statistic (3) the manipulations applied in [32] did not lead to the expected differences in the measurement of the three different cognitive load types [37]. In summary, their psychometric instrument might have measured only the level of expectation instead of the actual invested effort devoted in the complexity of the activity (intrinsic load), its ineffective explanations (extraneous) and its understanding (germane). To evaluate a more direct relation between the three types of load and the learning outcomes, a randomized experiment was performed, with bachelor students who received a description of the Bayes theorem. To measure the three different types of load, the authors changed the order of the rating scales and added three items to it (as per Table 7). These items were supposed to contribute to the evaluation of the internal consistency of the theoretical assumption that the three types of load are separated, additive and independent. Findings suggest that two items improved the internal consistency of the mental effort for intrinsic and extraneous loads but not for germane load, suggesting its re-definition [6].

Table 7. Informed subjective rating questionnaire proposed in [39] available to learners while rating different learning scenarios on a 7-point Likert scale (1 = very low, 7 = very high)

Type of load	Scale
Intrinsic load	I invested a very high mental effort in the complexity of this activity
Extraneous load	I invested a very high mental effort in unclear and ineffective explanations and instructions in this activity
Germane load	I invested a very high mental effort during this activity in enhancing my knowledge and understanding

Zukic and colleagues [38] focused on the assessment of the validity of the instrument developed in [37] as well as its internal consistency and its capability to correlate with learning outcomes. In their study, the correlations between intrinsic and extraneous load and between extraneous and germane load were statistically significant. A low degree of experienced intrinsic load and a high degree of reported germane load could explain the improvement of the learning outcomes. Additionally, a regression analysis verified that the items associated to the germane load could actually explain the perceived learning. Eventually, a confirmatory factor analysis supported the development of a three-dimensional model that includes the three types of load. The main take away of this study is that germane load can be measured as an independent source of load.

Klepsch et al. proposed an alternative way to measure the three load types reliably and validly [39]. The novelty of their approach is the use of two forms of ratings: informed and naïve. According to this, they conducted an experiment with two different group of learners. The first, the informed rating group, was trained on how to differentiate the three types of load through a theoretical explanation of CLT and its assumptions. The second, the naïve rating group, did not receive the training on CLT.

Learners were asked to rate 24 learning scenarios grouped in 5 different domains (language learning, biology, mathematics, technology and didactics). To detect changes

Table 8. Informed subjective rating questionnaire proposed in [39] available to learners while rating different learning scenarios on a 7-point Likert scale (1 = very low, 7 = very high)

Type of load	Scale
Intrinsic load	During this task, Intrinsic Load was…
Extraneous load	During this task, Extraneous Load was…
Germane load	During this task, Germane Load was…

in the cognitive load experienced by the two groups of learners, only one type of cognitive load was manipulated at a time. The learners in group one received the questionnaire in Table 8, while those in group two received the questionnaire in Table 9 Both the groups received also an additional question on perceived overall cognitive load adapted from [36]. The participants in the informed ratings group correctly discriminated intrinsic, extraneous and germane loads in line with the expectations. However, participants, in the Naïve ratings group, correctly discriminated only the intrinsic and the extraneous loads but they were not able to differentiate germane load.

Table 9. First version of the Naïve rating scales questionnaire proposed in [39] 7-point Likert scale (1 = completely wrong, 7 = absolutely right)

Type of load	Scale
Intrinsic load	• For this task, many things needed to be kept in mind simultaneously • This task was very complex
Germane load	• For this task, I had to highly engage myself • For this task, I had to think intensively what things meant
Extraneous load	• During this task, it was exhausting to find the important information • The design of this task was very inconvenient for learning • During this task, it was difficult to recognize and link the crucial information

A reliability analysis of the scales was executed, by task, using the Cronbach alpha measure based on the formula presented in [40]. This allowed to compute the mean of several given alpha values based on sampling distribution. The validity of the measure was analysed by comparing the ratings of learners with the expectations for each type of load for each task. A very low reliability was detected for all the tasks, in the informed ratings group, this being an indicator of the capability of learners to differentiate the types of load separately. However, in the naïve ratings groups, reliability was high, suggesting how the three types of load were not clearly separable. In particular, germane load was the dimension that was not discriminable across the two groups. Starting from this unsatisfying finding, the authors developed a new scale for germane load (Table 10).

Table 10. Second version of the Naïve rating questionnaire proposed in [39] with a new scale for the germane load on a 7-point Likert scale (1 = completely wrong, 7 = absolutely right)

Load type	Scale
Germane load	• I made an effort, not only to understand several details, but to understand the overall context • My point while dealing with the task was to understand everything correctly • The learning task consisted of elements supporting my comprehension of the task

Subsequently, they evaluated the overall new questionnaire with a larger sample. A new experiment was conducted with a group of students who received 8 tasks, one at a time, designed to induce more or less germane load. Here, in contrast to the first study, and in line with the idea of doing experiments in more realistic learning environments, each learning task was designed to induce changes in the three types of load. For intrinsic load, the degree of interactivity of the tasks was manipulated. For extraneous load, different learning formats were considered, some employing text and pictures together, some individually and some with additional non-relevant information. Eventually, germane load was manipulated by creating tasks aimed at eliciting different degrees of deeper learning processes. A reliability and validity analysis, conducted as in the first experiment, confirmed that it is possible to measure the three types of load separately, in line with the triarchic theory of load [2].

3.2 Task Performance and Self-reported Measures

Deleeuw and Mayer tested the separability of the three types of load in a multimedia lesson on the topic of electric motors [41]. Two experiments were executed: one with a pre-question on the content of the lesson aimed at motivating learners to focus on deeper cognitive processing, and one without. Authors manipulated extraneous load providing redundant instructional designs to learners. Similarly, they manipulated intrinsic load through changes on the complexity of the sentences that explained the lesson. Eventually, they examined the differences in the germane load by comparing students with high scores on a test of problem solving transfer, against students with lower scores. The authors evaluated the sensitivity of the response time to a secondary task during learning for measuring the extraneous load, the effort ratings during learning for measuring the intrinsic load, and the self-reported difficulty rating, after learning, for measuring the germane load (as per Table 11). In details, the secondary task consisted of a visual monitoring task where learners had to recognise a periodic change of colour and to press the space bar each time this colour change took place.

Table 11. Subjective rating scales and secondary task reaction time proposed by [41] on a 9-point Likert Scale (1 = extremely low, 9 = extremely high for intrinsic load and 1 = extremely easy, 9 = extremely difficult for germane load)

Types of load	Scale/measure
Intrinsic load	Your level of mental effort on this part of the lesson
Germane load	How difficult this lesson was
+	
Extraneous load	Measured by the response time to a secondary task At each of eight points in an animated narration, the background colour slowly changes (pink to black) Learner is required to press the spacebar of the computer as soon as the color changes

The findings of the experiment supported the triarchic theory of cognitive load [2]. Students who received redundant information needed longer reaction time than students who did not receive redundant instructional design. The explanation about the electric motor has been provided by learners using different sentences with different levels of complexity. The scale for intrinsic load reflected higher effort for high complexity sentences and lower effort for low complexity sentences. Students who reported a lower and a higher transfer reflected their difficulty by the rating scale provided: low transfer reflected high difficulty, high transfer reflected low difficulty. Thus, the authors showed that these different measures of load (reaction time, effort and difficulty) are sensitive to different types of load (extraneous, intrinsic, germane) [41]. The three different variables analysed (redundancy and complexity of statement, high or slow capacity of transfer to solve a problem) are strongly correlated with the three different types of load, thus providing evidence for their good sensitivity. Eventually, authors recommended a replication of their research study in other contexts and with different students because the measurement of the three cognitive load types might be often intrusive, creating an artificial learning situation. In addition, the study did not account for the prior knowledge of learners, (most of them had a low prior knowledge) as an important variable that could influence the overall perception of load.

Cerniak et al. [34] hypothesised how the split attention effect, proposed by Sweller [2], could be mediated not only by a reduction of extraneous cognitive load but also by an increase of germane load (germane load explanation) [42]. An experiment, conducted in a learning context on physiological processes of a nephron, was aimed at testing the above the research hypothesis. Authors employed the reaction time on a secondary as a task performance measure, in order to detect variations in the overall cognitive load between learners who received an integrated format of instructional designs and learners who received a split source format. The former learners were expected to experience less overall load because the integrated format was believed to decrease their extraneous load. The latter learners were expected to experience more overall load due to the split attention effect believed to increase their extraneous load, as suggested in [43]. In the experiment, learners had to press the space bar of the keyboard of a computer every time a stimulus appeared on the screen (for example the change of a colour). The longer time required to react to this secondary task, the higher cognitive load exerted on the primary task. Eventually, subjective ratings were applied to measure the three types of load, as per Table 12.

Table 12. Subjective rating scales for the cognitive load types proposed by [34] on a 6-point Likert Scale (1 = not at all, 6 = extremely)

Type of load	Scale
Intrinsic load	How much difficult was the learning content for you?
Extraneous load	How difficult was it for you to learn with the material?
Germane load	How much did you concentrate during learning?

Findings showed that there is no difference in the overall cognitive load between learners who received the split source format and those who received the integrated source format. As a consequence, the former learners increased their extraneous load and decreased their germane load, whereas the latter learners decreased their extraneous load and increased their germane load. This confirms that the extraneous and germane loads partially mediate the split attention effect. However, authors brought forward a critique whereby there could be a possible confusion between the two different questions designed for intrinsic and extraneous loads. Learners could have the impression to answer the same question. In fact, in a new learning context, learners might not be able to identify the source of difficulty that means the content or the instructional material delivered. The authors spotted a high correlation between the extraneous and the germane loads through an analysis of the learning outcomes. However, they did not state that these measures of loads were aimed at tackling different working memory resources. As a consequence, the relation between learning processes and working memory capacity was not demonstrated.

4 Synthesis and Observations on the Scientific Value of Cognitive Load Theory

According to the literature review conducted in the previous sections, it appears evident that the three types of load envisioned in Cognitive Load Theory – intrinsic, extraneous and germane – have been mainly measured by means of subjective rating scales. This has more a practical explanation because self-reporting scales are easier to use and they do not influence the primary task when compared to secondary task measures. They can be administered post-learning tasks and they are aimed at representing a perceptual subjective experience of a learner for an entire learning session. This is in contrast to secondary task measures which, even if more sensitive to variations of cognitive load, they are more intrusive since they alter the natural execution of a learning task. A number of researchers brought forward critiques on Cognitive Load Theory in relation to its theoretical clarity [44, 45] and its methodological approach [46]. According to these critiques, the assumptions of CLT appear circular because its three types of load are believed not to be empirically measurable. Empirical research is based on observed and recorded data and it derives knowledge from actual experience rather than from a theory, a belief or a logic coming from first principles 'a priori'. This is the case of subjective rating scales aimed at measuring the cognitive load types. Regardless of the way these scales relate to the evaluation of the different cognitive load types, all of them underlie the phenomenon they are pretending to measure in their premises or suppositions, namely the definitions of intrinsic, extraneous and germane loads (Fig. 7, left). In other words, the premises of CLT – its cognitive load types – are believed to be confirmed by the data coming from their measurements circularly, without empirical evidence.

In addition, the fact that human cognitive processes, related to the same instructional design, can be regarded as germane load in one case and as extraneous load in another case, it means that CLT can account for nearly every situation [45]. This critique also refers to the 'Expertise Reversal Effect' [47]. In fact, on one hand, some instructional design, such as written explanations followed by a graphic element to enhance its understanding, can be useful for a novice learner, by reducing the extraneous load and increasing the germane load. On the other hand, the same graphical aid can be useless for an expert learner because it can reduce germane load and increase extraneous load. In fact, for an expert, it can be redundant to read instructional designs just registered and automatized in own memory, hampering understanding and learning. Depending on the degree of expertise, the same instructional design can lead to germane or extraneous load, emphasising the circularity of CLT (Fig. 7, right). The theoretical differences regarding the types of load are based on the subjective experienced load of learners, implying that they are able to differentiate them by their own. This issue, as discussed above, depends on the way the questions are formulated, and on the familiarity of learners on the different cognitive load types, and their prior knowledge. All these variables are not easy to monitor and control, they can create confusion on the source of the supposed experienced load. Under a strict scientific view, the evaluation of this supposed load does not come from the experience of the learners, rather from the principles of the theory is based upon.

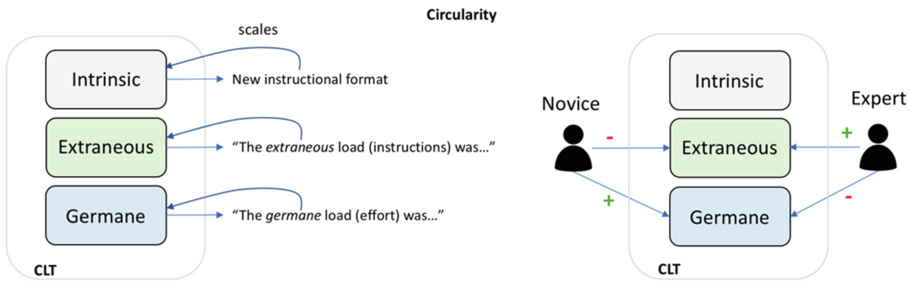


Fig. 7. The circularity of the load types of Cognitive Load Theory (left) and the 'expertise reversal effect' by which different cognitive processes can be regarded differently (right)

To analyse the scientific value of CLT, two different methodological approaches have been followed: the rationalism of Karl Popper [48, 49] and the structuralist approach of the theories of Joseph Sneed [50]. Under the former approach, it is not possible to consider CLT scientific because its basic principles, namely the three different types of load, cannot be tested by means of any experimental method, consequently they are not falsifiable [46] (Fig. 8, left). To be scientific, the measures should be sensitive to the different types of load. From a strict rationalist point of view, a measure is scientific if it does not presuppose the assumptions that it shall measure in its rationale [46]. However, as previously discussed, most of the subjective rating scales, conceived for the cognitive load types, contain the variables they pretend to measure. This implies that the logic of the questions influences the logic of the answers. In turn, the measures of the loads can be obtained 'a priori', by setting the questions to validate the theory they are pretending to verify, and not through any authentic experience of cognitive load. CLT should provide empirical evidence about the cognitive load types. Unfortunately, this has not convincingly emerged in the literature of

Educational Psychology, Instructional Design and Cognitive Load Theory, justifying the scepticism regarding the possibility to measure the three different types of load.

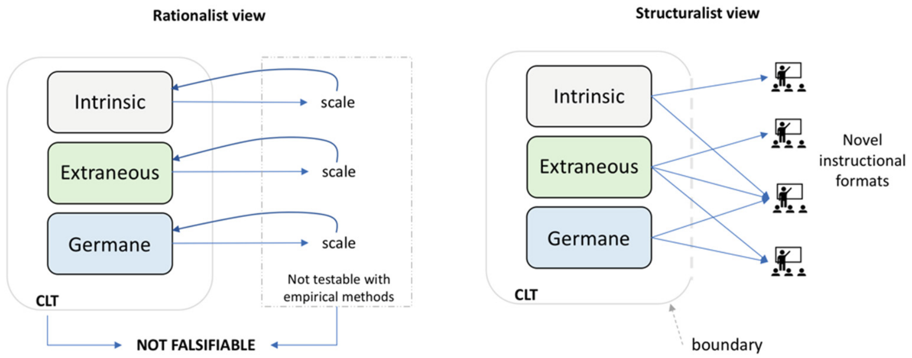


Fig. 8. The rationalist view of Cognitive Load Theory (left) and the structuralist view (right)

The second methodological approach to analyse CLT is based upon structuralism [50–53]. Under its logic, the scientific value of the theoretical principles of CLT does not depend on their empirical validity. Rather it depends upon their effectiveness to form the ground of the structure of a theory that consents to derive specific predictions on how detailed instructional manipulations can affect learning outcomes [46]. The structuralist analysis considers the fundamental assumptions of CLT as theoretical axioms. The empirical content of these axioms is valid in the context of the theory if they contribute to expand the theory itself (Fig. 8, right). Regardless whether it is possible to validate some research predictions or not, these predictions can still expand the theory. In fact, CLT has been extensively adopted for the design of several new instructional formats, expanding its boundaries [46]. As discussed in the previous sections, several research experiments have been performed in different learning contexts. In each of this, the intrinsic, extraneous and germane loads have been manipulated, individually or in pair by employing the traditional experimental/control group design. In turn, the cognitive load of learners and their learning outcomes where analysed [6]. If this analysis showed that learning has been actually facilitated, and statistical power held, then it means that a new instructional design was conceived as it actually promoted one or more types of load. Similarly, starting from the study of the 'Goal free effects' compared to the traditional ways to solve a problem (means analysis), Sweller and his colleagues have produced various novel findings and approaches to inform instructional design. Yet, Plass et al. [54] provided a complete list of CLT effects such as the 'Worked completion effect' [55], the 'Split attention effect' [56] the 'Redundancy effect' [57], the 'Modality effect' [58], the 'Expertise reversal effect' [47] and the 'Collective working memory effect' [11]. As a consequence, according to a structuralist point of view, Sweller stated that CLT has been developed and evolved as a consequence of these contributions and experiments [6]. They defend the fact that the three types of load were not elaborated a priori, rather they have been developed according to experimental findings that are falsifiable in their nature. In fact, it is still possible to replicate the experiments and obtain opposite findings. However, what cannot be considered falsifiable is only the definition of the three types of load employed in different experiments because the measures adopted are not considered scientific. In short, Sweller and colleagues strongly support the view that CLT is actually built upon empiricism [59]. As educational psychologist, Sweller and Chandler [60] share the same ultimate goal in the context of cognition and instruction: the generation of new, helpful instructional techniques aimed at improving learning.

5 Reconceptualization of Cognitive Load Types

As a consequence of the critiques related to the theoretical development of CLT and after several failed attempts to find a generally applicable measurement technique as well as the development of measures for the three different types of load, the theory has been re-conceptualised using the notion of element interactivity. This refers to the numbers of elements that must be processed simultaneously in working memory for schema construction and their interactions [18]. In this update of CLT, the element interactivity now defines the mechanisms not only of intrinsic load, but also of extraneous load [6]. In detail, the extraneous load is related to the degree of interactivity of the elements of the instructional material used for teaching activities, and instructional designs should be aligned to this. These designs should not focus on enhancing the number of items to be processed by learners, otherwise the resulting load could be considered extraneous. In other words, when instructional designs do not add instructions that increase the number of elements that must be processed within working memory, then the germane load of learners can be triggered. In this case, existing instructions can facilitate the use of working memory allocated for the intrinsic load. Additionally, germane load is no longer an independent source of load, it is a function of those working memory resources related to the intrinsic load of the task. In turn, intrinsic load depends on the characteristic of the task, extraneous load on the characteristic of the instructional material, on the characteristic of the instructional design and on the prior knowledge of learners. Eventually, germane load depends on the characteristics of a learner which equates to the resources of working memory allocated to deal with the intrinsic load [6] (Fig. 9).

The main theoretical contradiction before the reconceptualization of CLT was the additivity and the compensability of germane and extraneous loads. Here, the critical point is that, if extraneous load decreases, while keeping intrinsic load constant, then germane load should increase too. However, the measures for the three cognitive loads appeared in the last 30 years, confirm that this compensation does not have empirical evidence: the total cognitive load does not remain constant but changes [6]. After the reconceptualization (Fig. 9), germane load is related to that part of working memory that deals with the degree of element interactivity of the task. It can be promoted by creating instructional design aligned to it but it also depends on the intrinsic load, and as a consequence, it is not clearly measurable [6]. In fact, germane load now forms a balanced whole with extraneous load without creating logical and empirical contradictions. If intrinsic load remains constant but extraneous load changes, the overall

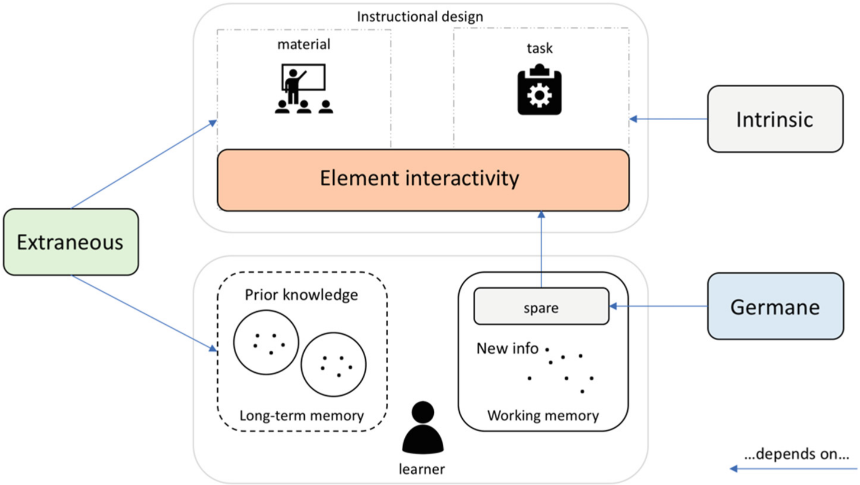


Fig. 9. Redefinition of the cognitive load types and their roles.

cognitive load changes too because more or less working memory resources are devoted to deal with the degree of element interactivity. At a given level of knowledge and expertise, intrinsic load cannot be altered without changing the content of the material presented to learners altogether. Extraneous load, instead, can be altered by changing the instructional procedures. Yet, germane load coincides to those working memory resources allocated to deal with the degree of element interactivity inherent to an underlining learning task. Although germane load has now a fundamental role to deal with intrinsic load, the additivity of CLT still holds in the two remaining theoretical assumptions: the intrinsic and the extraneous load. According to this reconceptualisation, most of the critiques related to the circularity of CLT do not longer stand according to Sweller. Additionally, Sweller and colleagues, as in [59], consider the unidimensional subjective rating scale of mental effort proposed by Paas et al. [13] a valid measure of overall cognitive load. In fact, if intrinsic load is kept constant, it is feasible to measure the extraneous load by only altering the instructional designs between an experimental and a control group. It is also possible to measure one type of load keeping the other constant, and the overall load measured would be an indicator of the modified type of load: extraneous or intrinsic.

6 Final Remarks

The measurement of the cognitive load types envisioned in the Cognitive Load Theory is a critical challenge for its theoretical development and its scientific value. After the literature review conducted in the previous sections, and after the presentation of the critiques that brought to the reconceptualization of the cognitive load types, the reader is left with two possible interpretations. On one hand, germane load is not clearly measurable by a common and standardised way, consequently its theoretical independence is denied, and after its reconceptualisation, it is a function of intrinsic load. On the other hand, there is evidence, triggered by the proposal of novel multiple subjective rating scales [38, 39], that the three types of load are measurable, even the most challenging, namely the germane load. Sweller and colleagues believe that germane load exists but it is not measurable [59]. He suggested that one of the most reliable way to measure the overall cognitive load is the unidimensional subjective scale of Mental Effort [13]. However, the fact that unidimensional scale has been widely employed within Educational Psychology, does not mean it is always the most appropriate. For example, within the discipline of Human Factors (Ergonomics), there exist a plethora of empirical studies that all point to the multi-dimensional nature of the construct of Mental Workload (cognitive load for educational psychologists) [16, 61–65]. There is also an emerging body of knowledge, within Computer Science, that is employing more formal non-linear approaches for modelling mental workload as a multi-dimensional construct [66–68]. Similarly, applications of mental workload as a multi-dimensional concept can be found in Human Computer Interaction [69–72].

Learning is a complex process, it is hard to evaluate it mostly because it is perceived as a subjective one. Similarly, cognitive load it is a complex construct and it is assumed it can be modelled end evaluated through quantitative criteria to satisfy the empirical exigencies of scientific research. This is an existing methodological gap and it is the reason why, so far, there is little evidence of generally applicable subjective measurement techniques and measures for the three types of load and for overall cognitive load. According to Klepsch and colleagues, their informed rating scale is a novelty in CLT research and it seems to be a valid method for measuring the three types of load [39]. They believe it is the most logical approach because if learners are informed, then the evaluation of the experienced load can be done with a higher degree of awareness. However, in our view, this might bring back the issue of circularity, suggesting that we are leading learners to understand Cognitive Load Theory as well as its assumptions and influence them to rate their subjective experience to fit our expectations. Cognitive load is a complex construct and indeed CLT has had a significant impact for instructional design. Circularity is also an important issue that should be avoided in favour of empiricism and falsifiability of measures. We believe that, with advances in technologies and the availability of cheap sensors and non-invasive instruments for gathering responses of the human brain and bodies, physiological measures of mental workload might finally shed further light on the complex but fascinating problem of cognitive load modelling.

References

1. Sweller, J., Van Merrienboer, J.J., Paas, F.G.: Cognitive architecture and instructional design. Educ. Psychol. Rev. **10**(3), 251–296 (1998)
2. Young, M.S., Brookhuis, K.A., Wickens, C.D., Hancock, P.A.: State of science: mental workload in ergonomics. Ergonomics **58**(1), 1–17 (2015). https://doi.org/10.1080/00140139.2014.956151
3. Baddeley, A.: Recent developments in working memory. Curr. Opin. Neurobiol. **8**(2), 234–238 (1998). https://doi.org/10.1016/s0959-4388(98)80145-1

4. Paivio, A.: Mental Representation: A Dual-Coding Approach. Oxford University Press, New York (1986)
5. Hart, S.G., Wickens, C.D.: Workload assessment and prediction. In: Booher, H.R. (ed.) MANPRINT: An Approach to Systems Integration, pp. 257–296. Springer, Dordrecht (2012). https://doi.org/10.1007/978-94-009-0437-8_9
6. Sweller, J.: Element interactivity and intrinsic, extraneous, and germane cognitive load. Educ. Psychol. Rev. **22**(2), 123–138 (2010)
7. Miller, G.A.: The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychol. Rev. **63**(2), 81 (1956)
8. Clark, R.C., Nguyen, F., Sweller, J., Baddeley, M.: Efficiency in learning: evidence-based guidelines to manage cognitive load. Perform. Improv. **45**(9), 46–47 (2006)
9. Mayer, R.E., Moreno, R.: Nine ways to reduce cognitive load in multimedia learning. Educ. Psychol. **38**(1), 43–52 (2003)
10. Mayer, R.E.: Principles for managing essential processing in multimedia learning: segmenting, pretraining, and modality principles. In: The Cambridge Handbook of Multimedia Learning, pp. 169–182 (2005)
11. Kirschner, F., Paas, F., Kirschner, P.A.: Superiority of collaborative learning with complex tasks: a research note on an alternative affective explanation. Comput. Hum. Behav. **27**(1), 53–57 (2011)
12. Chi, M.T.H., Glaser, R., Rees, E.: Expertise in problem solving. In: Sternberg, R. (ed.) Advances in the Psychology of Human Intelligence, pp. 7–75. Lawrence Erlbaum, Hillsdale (1982)
13. Paas, F.G., Van Merriënboer, J.J., Adam, J.J.: Measurement of cognitive load in instructional research. Percept. Mot. Skills **79**(1), 419–430 (1994)
14. Longo, L., Barrett, S.: A computational analysis of cognitive effort. In: Nguyen, N.T., Le, M. T., Świątek, J. (eds.) ACIIDS 2010, Part II. LNCS (LNAI), vol. 5991, pp. 65–74. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-12101-2_8
15. Longo, L., Barrett, S.: Cognitive effort for multi-agent systems. In: Yao, Y., Sun, R., Poggio, T., Liu, J., Zhong, N., Huang, J. (eds.) BI 2010. LNCS (LNAI), vol. 6334, pp. 55–66. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15314-3_6
16. Longo, L., Leva, M.C.: Human Mental Workload: Models and Applications, vol. 726. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61061-0
17. Halford, G.S., Maybery, M.T., Bain, J.D.: Capacity limitations in children's reasoning: a dual-task approach. Child Dev. **57**(3), 616–627 (1986). https://doi.org/10.2307/1130340
18. Sweller, J.: Cognitive load theory, learning difficulty, and instructional design. Learn. Instr. **4** (4), 295–312 (1994)
19. Paas, F.G., Van Merriënboer, J.J.: Variability of worked examples and transfer of geometrical problem-solving skills: a cognitive-load approach. J. Educ. Psychol. **86**(1), 122 (1994)
20. Gerjets, P., Scheiter, K.: Goal configurations and processing strategies as moderators between instructional design and cognitive load: evidence from hypertext-based instruction. Educ. Psychol. **38**(1), 33–41 (2003)
21. Miwa, K., Kojima, K., Terai, H., Mizuno, Y.: Measuring cognitive loads based on the mental chronometry paradigm. In: COGNITIVE 2016 : The Eighth International Conference on Advanced Cognitive Technologies and Applications, pp. 38–41 (2016)
22. Meyer, D.E., Osman, A.M., Irwin, D.E., Yantis, S.: Modern mental chronometry. Biol. Psychol. **26**(1–3), 3–67 (1988)

23. Gerjets, P., Scheiter, K., Opfermann, M., Hesse, F.W., Eysink, T.H.S.: Learning with hypermedia: the influence of representational formats and different levels of learner control on performance and learning behavior. Comput. Hum. Behav. **25**(2), 360–370 (2009). https://doi.org/10.1016/j.chb.2008.12.015
24. Corbalan, G., Kester, L., van Merriënboer, J.J.G.: Selecting learning tasks: effects of adaptation and shared control on learning efficiency and task involvement. Contemp. Educ. Psychol. **33**(4), 733–756 (2008). https://doi.org/10.1016/j.cedpsych.2008.02.003
25. Ayres, P.: Using subjective measures to detect variations of intrinsic cognitive load within problems. Learn. Instr. **16**(5), 389–400 (2006)
26. Gerjets, P., Scheiter, K., Catrambone, R.: Can learning from molar and modular worked examples be enhanced by providing instructional explanations and prompting self-explanations? Learn. Instr. **16**(2), 104–121 (2006). https://doi.org/10.1016/j.learninstruc.2006.02.007
27. Hart, S.G., Staveland, L.E.: Development of NASA-TLX (task load index): results of empirical and theoretical research. In: Advances in Psychology, vol. 52, pp. 139–183. Elsevier (1988). https://doi.org/10.1016/s0166-4115(08)62386-9
28. Scheiter, K., Gerjets, P., Catrambone, R.: Making the abstract concrete: visualizing mathematical solution procedures. Comput. Hum. Behav. **22**(1), 9–25 (2006). https://doi.org/10.1016/j.chb.2005.01.009
29. Gerjets, P., Scheiter, K., Catrambone, R.: Designing instructional examples to reduce intrinsic cognitive load: molar versus modular presentation of solution procedures. Instr. Sci. **32**(1/2), 33–58 (2004). https://doi.org/10.1023/b:truc.0000021809.10236.71
30. Galy, E., Cariou, M., Mélan, C.: What is the relationship between mental workload factors and cognitive load types? Int. J. Psychophysiol. **83**(3), 269–275 (2012). https://doi.org/10.1016/j.ijpsycho.2011.09.023
31. Thayer, R.E.: Toward a psychological theory of multidimensional activation (arousal). Motiv. Emot. **2**(1), 1–34 (1978)
32. Leppink, J., Paas, F., Van der Vleuten, C.P.M., Van Gog, T., Van Merriënboer, J.J.G.: Development of an instrument for measuring different types of cognitive load. Behav. Res. Methods **45**(4), 1058–1072 (2013). https://doi.org/10.3758/s13428-013-0334-1
33. Kalyuga, S., Chandler, P., Sweller, J.: Levels of expertise and instructional design. Hum. Factors **40**(1), 1 (1998)
34. Cierniak, G., Scheiter, K., Gerjets, P.: Explaining the split-attention effect: Is the reduction of extraneous cognitive load accompanied by an increase in germane cognitive load? Comput. Hum. Behav. **25**(2), 315–324 (2009). https://doi.org/10.1016/j.chb.2008.12.020
35. Salomon, G.: Television is “easy” and print is “tough”: the differential investment of mental effort in learning as a function of perceptions and attributions. J. Educ. Psychol. **76**(4), 647 (1984)
36. Paas, F.G.: Training strategies for attaining transfer of problem-solving skill in statistics: a cognitive-load approach. J. Educ. Psychol. **84**(4), 429 (1992)
37. Leppink, J., Paas, F., van Gog, T., van der Vleuten, C.P.M., van Merriënboer, J.J.G.: Effects of pairs of problems and examples on task performance and different types of cognitive load. Learn. Instr. **30**, 32–42 (2014). https://doi.org/10.1016/j.learninstruc.2013.12.001
38. Zukić, M., Đapo, N., Husremović, D.: Construct and predictive validity of an instrument for measuring intrinsic, extraneous and germane cognitive load. Univ. J. Psychol. **4**(5), 242–248 (2016). https://doi.org/10.13189/ujp.2016.040505
39. Klepsch, M., Schmitz, F., Seufert, T.: Development and validation of two instruments measuring intrinsic, extraneous, and germane cognitive load. Front. Psychol. **8** (2017). https://doi.org/10.3389/fpsyg.2017.01997

40. Rodriguez, M.C., Maeda, Y.: Meta-analysis of coefficient alpha. Psychol. Methods **11**(3), 306 (2006)
41. DeLeeuw, K.E., Mayer, R.E.: A comparison of three measures of cognitive load: Evidence for separable measures of intrinsic, extraneous, and germane load. J. Educ. Psychol. **100**(1), 223–234 (2008). https://doi.org/10.1037/0022-0663.100.1.223
42. Kester, L., Kirschner, P.A., van Merriënboer, J.J.G.: The management of cognitive load during complex cognitive skill acquisition by means of computer-simulated problem solving. Br. J. Educ. Psychol. **75**(1), 71–85 (2005). https://doi.org/10.1348/000709904x19254
43. Van Gog, T., Paas, F.: Instructional efficiency: revisiting the original construct in educational research. Educ. Psychol. **43**(1), 16–26 (2008)
44. Schnotz, W., Kürschner, C.: A reconsideration of cognitive load theory. Educ. Psychol. Rev. **19**(4), 469–508 (2007)
45. De Jong, T.: Cognitive load theory, educational research, and instructional design: some food for thought. Instr. Sci. **38**(2), 105–134 (2010)
46. Gerjets, P., Scheiter, K., Cierniak, G.: The scientific value of cognitive load theory: a research agenda based on the structuralist view of theories. Educ. Psychol. Rev. **21**(1), 43–54 (2009). https://doi.org/10.1007/s10648-008-9096-1
47. Sweller, J., Ayres, P.L., Kalyuga, S., Chandler, P.: The expertise reversal effect. The expertise reversal effect. Educ. Psychol. **38**(1), 23–31 (2003)
48. Popper, K.: The Logic of Scientific Discovery. Routledge, Abingdon (2005)
49. Popper, K.: Conjectures and Refutations: The Growth of Scientific Knowledge. Routledge, Abingdon (2014)
50. Sneed, J.D.: The Logical Structure of Mathematical Physics, vol. 3. Springer, Heidelberg (2012). https://doi.org/10.1007/978-94-009-9522-2
51. Balzer, W., Moulines, C., Sneed, J.: An architectonic for science–the structuralist program. Reidel, Dordrecht (1987). Springer. Traducción al castellano de P. Lorenzano: Una arquitectónica para la ciencia–El programa estructuralista. Universidad Nacional de Quilmes, Bernal (Argentina) (2012)
52. Stegmüller, W.: The Structuralist View of Theories: A Possible Analogue of the Bourbaki Programme in Physical Science. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-95360-6
53. Westermann, R.: Structuralist reconstruction of psychological research: cognitive dissonance. Ger. J. Psychol. **12**(3), 218–231 (1988)
54. Plass, J.L., Moreno, R., Brünken, R.: Cognitive Load Theory. Cambridge University Press, Cambridge (2010)
55. Sweller, J., Cooper, G.A.: The use of worked examples as a substitute for problem solving in learning algebra. Cogn. Instr. **2**(1), 59–89 (1985)
56. Tarmizi, R.A., Sweller, J.: Guidance during mathematical problem solving. J. Educ. Psychol. **80**(4), 424 (1988)
57. Chandler, P., Sweller, J.: Cognitive load theory and the format of instruction. Cogn. Instr. **8**(4), 293–332 (1991). https://doi.org/10.1207/s1532690xci0804_2
58. Mousavi, S.Y., Low, R., Sweller, J.: Reducing cognitive load by mixing auditory and visual presentation modes. J. Educ. Psychol. **87**(2), 319 (1995)
59. Sweller, J., Ayres, P., Kalyuga, S.: Cognitive Load Theory, Explorations in the Learning Sciences, Instructional Systems, and Performance Technologies 1. Springer, New York (2011). https://doi.org/10.1007/978-1-4419-8126-4
60. Sweller, J., Chandler, P.: Evidence for cognitive load theory. Cogn. Instr. **8**(4), 351–362 (1991)

61. Longo, L.: A defeasible reasoning framework for human mental workload representation and assessment. Behav. Inf. Technol. **34**(8), 758–786 (2015)
62. Longo, L.: Formalising human mental workload as a defeasible computational concept. Ph. D. thesis, Trinity College Dublin (2014)
63. Longo, L.: Mental workload in medicine: foundations, applications, open problems, challenges and future perspectives. In: 2016 IEEE 29th International Symposium on Computer-Based Medical Systems (CBMS), pp. 106–111 (2016). https://doi.org/10.1109/cbms.2016.36
64. Wickens, C.D.: Mental workload: assessment, prediction and consequences. In: Longo, L., Leva, M.C. (eds.) H-WORKLOAD 2017. CCIS, vol. 726, pp. 18–29. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61061-0_2
65. Hancock, P.A.: Whither workload? Mapping a path for its future development. In: Longo, L., Leva, M.C. (eds.) H-WORKLOAD 2017. CCIS, vol. 726, pp. 3–17. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61061-0_1
66. Rizzo, L., Dondio, P., Delany, S.J., Longo, L.: Modeling mental workload via rule-based expert system: a comparison with NASA-TLX and workload profile. In: Iliadis, L., Maglogiannis, I. (eds.) AIAI 2016. IAICT, vol. 475, pp. 215–229. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-44944-9_19
67. Moustafa, K., Luz, S., Longo, L.: Assessment of mental workload: a comparison of machine learning methods and subjective assessment techniques. In: Longo, L., Leva, M.C. (eds.) H-WORKLOAD 2017. CCIS, vol. 726, pp. 30–50. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61061-0_3
68. Rizzo, L.M., Longo, L.: Representing and inferring mental workload via defeasible reasoning: a comparison with the NASA Task Load Index and the Workload Profile. In: Proceedings of the 1st Workshop on Advances in Argumentation in Artificial Intelligence co-located with {XVI} International Conference of the Italian Association for Artificial Intelligence (AI*IA 2017), pp. 126–140 (2017)
69. Longo, L., Dondio, P.: On the relationship between perception of usability and subjective mental workload of web interfaces. In: IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, WI-IAT 2015, Singapore, 6–9 December, vol. I, pp. 345–352 (2015)
70. Longo, L.: Formalising human mental workload as non-monotonic concept for adaptive and personalised web-design. In: Masthoff, J., Mobasher, B., Desmarais, M.C., Nkambou, R. (eds.) UMAP 2012. LNCS, vol. 7379, pp. 369–373. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-31454-4_38
71. Longo, L.: Subjective usability, mental workload assessments and their impact on objective human performance. In: Bernhaupt, R., Dalvi, G., Joshi, A., Balkrishan, D.K., O'Neill, J., Winckler, M. (eds.) INTERACT 2017. LNCS, vol. 10514, pp. 202–223. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67684-5_13
72. Longo, L.: Experienced mental workload, perception of usability, their interaction and impact on task performance. PLoS ONE **13**(8), e0199661 (2018)

Developing an Objective Indicator of Fatigue: An Alternative Mobile Version of the Psychomotor Vigilance Task (m-PVT)

Michael Scott Evans[1(✉)], Daniel Harborne[2], and Andrew P. Smith[1]

[1] School of Psychology, Cardiff University, 63 Park Place, Cardiff, Wales CF10 3AS, UK
EvansMS3@cardiff.ac.uk
[2] School of Computer Science and Informatics, Cardiff University, Cardiff, Wales, UK

Abstract. Approximately 20% of the working population report symptoms of feeling fatigued at work. The aim of the study was to investigate whether an alternative mobile version of the 'gold standard' Psychomotor Vigilance Task (PVT) could be used to provide an objective indicator of fatigue in staff working in applied safety critical settings such as train driving, hospital staffs, emergency services, law enforcements, etc., using different mobile devices. 26 participants mean age 20 years completed a 25-min reaction time study using an alternative mobile version of the Psychomotor Vigilance Task (m-PVT) that was implemented on either an Apple iPhone 6s Plus or a Samsung Galaxy Tab 4. Participants attended two sessions: a morning and an afternoon session held on two consecutive days counterbalanced. It was found that the iPhone 6s Plus generated both mean speed responses (1/RTs) and mean reaction times (RTs) that were comparable to those observed in the literature while the Galaxy Tab 4 generated significantly lower 1/RTs and slower RTs than those found with the iPhone 6s Plus. Furthermore, it was also found that the iPhone 6s Plus was sensitive enough to detect lower mean speed of responses (1/RTs) and significantly slower mean reaction times (RTs) after 10-min on the m-PVT. In contrast, it was also found that the Galaxy Tab 4 generated mean number of lapses that were significant after 5-min on the m-PVT. These findings seem to indicate that the m-PVT could be used to provide an objective indicator of fatigue in staff working in applied safety critical settings such as train driving, hospital staffs, emergency services, law enforcements, etc.

Keywords: Psychomotor Vigilance Task (PVT) · Mental workload · Occupational fatigue · Objective indicator of fatigue · Attention

1 Introduction

In order to be able to meet task demands, there is usually a required amount of operator resources needed, referred to as human mental workload [1]. According to Hart and Staveland [2], human mental workload can be defined as a 'cost incurred by a human

L. Longo and M. C. Leva (Eds.): H-WORKLOAD 2018, CCIS 1012, pp. 49–71, 2019.
https://doi.org/10.1007/978-3-030-14273-5_4

operator to achieve a particular level of performance' and evolves from interactions between task demands, circumstances, skills, behaviour, and perceptions'. Therefore, human mental workload – often referred to as cognitive load – can be intuitively defined as the amount of mental work necessary for a person to complete a task over a given period of time [3, 4]. However, nowadays human mental workload is more generally defined as the measurement of the amount of mental resources involved in a cognitive task [5].

Human mental workload can be measured in real time using a variety of psychological and physiological techniques, which include; subjective psychological self-reported measures e.g., the NASA Task Load Index (NASA-TLX) [2, 6–8] and the NASA-MATB (National Aeronautics and Space Administration Multi-Attribute Task Battery [9] as well as objective physiological measures e.g., heart rate (HR), galvanic skin response (GSR), body temperature, electrocardiogram (ECG), electroencephalogram (EEG), and eye tracking [8, 10–19], and which have been extensively examined in various safety critical environments including; aviation [7, 20], train driving [21], car driving [22–24], and in an operating theater [6] but to name a few.

According to Wickens [25], the greatest value of conducting scientific human mental workload research is to be able to predict the consequences of high mental workload on performance. In other words, to better understand an individual's decision to consciously engage in a safe behaviour or in a potentially dangerous behaviour that could have devastating consequences. As a result, the concept of human mental workload has long been recognised as an important factor in individual performance [26–29]. Xie and Salvendy [29] state that both underload (i.e., low mental workload) and overload (i.e., high mental workload) degrade performance, whereby high and low levels of human mental workload have been shown to lead to operator error [22]. Longo [3] outlines that during low mental workload, individuals are more likely to experience levels of frustration and annoyance when processing information, which could result in an increase in their reaction time (RT). In contract, during high mental workload, individuals could experience confusion, which may result in a decrease in their information processing capacity, which could directly increase the likelihood of errors and mistakes. Therefore, these low and high mental workload information processing stages could have potentially dangerous consequences, especially in safety critical environments. Byrne [30] points out that the main application of mental workload has been to investigate situations where cognitive demand exceeds the acceptable safety tolerance threshold so that workload can be effectively reduced. Therefore, in high risk safety critical environments, the measurement of mental workload is of upmost importance due to its potential implications [31]. However, Xie and Salvendy [29] identified that the effect of fatigue on mental workload is not often considered in human mental workload research. Nevertheless, research carried out by Smith and Smith [32] on conductors/guards and engineers from the rail industry who work in high risk safety critical environments found that workload increased fatigue. However, subjective measures were predominately used in Smith and Smith's study. As a result, there is a need for an alternative mobile objective indicator of fatigue that can be used in high risk safety critical environments. In a controlled laboratory setting, the human Psychomotor Vigilance Task (PVT) [see 33, 34, for review] has become the widely accepted 'gold standard' tool for assessing the impact of fatigue on human cognitive neurobehavioral performance for monitoring temporal dynamic changes in attention [35–38]. The aim of the study was to investigate whether an alternative mobile version of the 'gold standard' Psychomotor Vigilance Task (PVT) could be used to provide an objective indicator of fatigue in staff working in applied safety critical settings, such as train driving, hospital staffs, emergency services, law enforcements, etc.

The rest of the paper is organised as follows. Section 2 describes related work on the Psychomotor Vigilance Task (PVT) while also extracting relevant studies to identify the gaps and rationale for the need of an alternative objective indicator of fatigue in staff working in applied safety critical settings. Section 3 outlines the design and empirical methodology of the proposed alternative mobile Psychomotor Vigilance Task (m-PVT). Section 4 presents the empirical results and discussion of the m-PVT. Finally, Sect. 5 provides a critical conclusion of the proposed alternative m-PVT and suggestions for future work.

2 Related Work

The Psychomotor Vigilance Task (PVT) can be traced back from the early work in simple reaction time (SRT) studies that were carried out by Wilhelm Maximilian Wundt (1832–1920) and continued by James McKeen Cattell (1860–1944) [39]. It is important to note that the modern PVT has been refined several times over the years [40–42] from its original development by Dinges and Powell [33] and has been shown to be sensitive to sleep deprivation, fatigue, drug use, and age. The PVT has also been widely implemented using a handheld device known as the PVT-192 (Ambulatory Monitoring Inc., Ardsley, New York, USA), as well as being extensively validated by various researchers [40, 43–47].

According to Basner, Mcguire, Goel, Rao and Dinges [48] and Basner et al. [49], the PVT-192 records participants' sustained attention based on repeated reaction time (RT) trials to visual stimuli that occur at random inter-stimulus intervals (ISI) that are between 2–10 s, for a standard 10-min period. In summary, the PVT-192 device operated by presenting participants with a stimulus that consisted of a four-digit millisecond counter that appears in a light-emitting diode (LED) dot-matrix display. The response consisted of a left or right button press, which depended on the configuration of the PVT-192 setup. The time difference between the stimulus presentation and the response constituted the participant's reaction time (RT). Each RT value was stored in the device and then uploaded to a personal computer, where the individual RTs are post-processed with the REACT software (Ambulatory Monitoring Inc., Ardsley, New York, USA), or other commercially available software, into summary statistics, such as the mean RT or the mean number of lapses (RTs $\geq$ 500 ms) per session [33, 40, 48, 50, 51]. For example, in Roach, Dawson, and Lamond's study [45], each participant performed either 5 min or 10 min RT sessions spaced at predetermined intervals (e.g., every 2 h) for a prolonged duration (e.g., 28 h), where each session consists of either 50 trials (equivalent to 5 min), or 100 trials (equivalent to 10 min). However, Khitrov et al. [52] tested the average delay of the PVT-192 and found that the recorded delay was greater than what was stated by the PVT-192 manufacturer. The delay recorded by the researchers was on average 2.4 ms greater when compared to the manufacturer's reported delay of 1 ms. Nevertheless, it is important to highlight that Khitrov et al. [52]

did acknowledge the possibility that the difference found could have been due to the non-instantaneous nature of the light detection circuit, or the actual delay associated with the PVT-192, since their experimental design did not permit them to be able to distinguish between these possibilities.

Dinges and Powell [33] have shown that the 10-min PVT is highly reliable. Roach, Dawson and Lamond [45] wanted to investigate whether 90 s could also be sufficiently sensitive enough to detect the effects of fatigue in comparison to their earlier research [see 43, for review], where they were able to find significant fatigue-related impairment during the first 5-min of a 10-min PVT. In this study, the researchers compared participants' neurobehavioral performance using the PVT between three different time durations (90 s, 5-min, and 10-min) to identify whether a shorter PVT could also be sensitive enough to detect the effects of fatigue. They found that it was only possible to implement a 5-min PVT as a substitute of the 10-min PVT, and not a 90 s PVT, thus only further supporting their earlier research [43]. However, it is important to note that analyses of their study were carried out using the mean RT and not the mean speed response (1/RT). Basner and Dinges [43] have identified that the mean RTs should not be the primary measure of alertness, and instead considering using the alternative primary measure of 1/RTs. In a later study, Basner, Mollicone and Dinges [42] aimed to further shorten the 5-min PVT [45] by developing a modified 3-min version of the PVT (PVT-B). They found that this 3-min version could be a useful tool for assessing behavioural alertness in settings where the 'gold standard' 10-min PVT could be more difficult or impractical to implement due to the nature of the study or location. However, further validation is required to determine whether both the 5-min PVT and PVT-B versions could indeed be sensitive enough to detect reduced levels of fatigue. Therefore, this study aimed to investigate a mobile version of the Psychomotor Vigilance Task (m-PVT) that could also be used to provide an objective indicator of fatigue in staff working in applied safety critical settings such as train driving, hospital staffs, emergency services, law enforcements, etc.

3 Design and Methodology

The aim of the study was to investigate whether an alternative mobile version of the 'gold standard' Psychomotor Vigilance Task (PVT) could be used to provide an objective indicator of fatigue in staff working in applied safety critical settings such as train driving, hospital staffs, emergency services, law enforcements, etc. The study received ethics approval from Cardiff University's Ethics Committee (EC.16.02.09.4464R). The study conformed to the seventh amendment of the Declaration of Helsinki 1964 [53] and all participants gave their informed written as well as electronic consent following the explanation of the nature of the study in written form.

3.1 Participants

26 (3 male and 23 female) participants with a mean age of 20 years (SD = 1.66) were recruited as volunteers from Cardiff University via the Experimental Management System (EMS) to take part in the study. The study involved participants attending two

sessions, a morning session (i.e., before 11:00) and an afternoon session (i.e., after 17:00), which were held on two consecutive days and counterbalanced, in exchange for £10. The study lasted 60 min in total for both sessions.

3.2 Materials/Apparatus

The mobile Psychomotor Vigilance Task (m-PVT) was presented to participants on one of two mobile devices: Apple's iPhone 6s Plus running Apple's iOS version 9.3.1 (Apple Inc.) or Samsung's Galaxy Tab 4 (Samsung Electronics Co. Ltd.) running on Android's operating system (OS) version 4.4.2 KitKat (Alphabet Inc.). The m-PVT ran in the following hardware configurations for the iPhone 6s Plus: system chip (Apple A9 APL1022), processor (Dual-core, 1840 MHz, Twister, 64-bit), graphics processor (PowerVR GT7600), and system memory (2048 MB RAM), and for the Samsung Galaxy Tab 4: system chip (Marvell PXA1088), processor (Quad-core, 1200 MHz, ARM Cortex-A7), graphics processor (Vivante), and system memory (1536 MB RAM). The iPhone 6s Plus had the following hardware configurations: the m-PVT was displayed on either a 5.5-in. (diagonal) 1920 × 1080-pixel native resolution at 401 ppi Retina high definition display (iPhone 6s Plus), or a 7-in. (diagonal) 1280 × 800-pixel (WXGA) native resolution at 216 pixels per inch (ppi) liquid crystal display (LCD) display (Samsung Galaxy Tab 4).

The m-PVT was programmed using the client code HTML, and CSS for the page visualisation and layout. JavaScript was also used to initiate the m-PVT, which was run using the Dolphin Web Browser (MoboTap Inc.) on both an Apple's iPhone 6s Plus and Samsung Galaxy Tab 4 (Dolphin Web Browser versions; Apple app version 9.9.0, and Android app version 11.5.6, respectively). The rationale for selecting the Dolphin Web Browser for this study was that it allowed the full screen feature to be enabled across the two different operating systems (OS), Apple iOS and Android OS platforms for both mobile devices. Other more native mobile internet browsers of each OS platform, such as Safari (Apple) and Chrome (Android) including Firefox, to name a few, did not permit full screen. Qualtrics Surveys (Qualtrics Labs, Inc.) were also used to collect demographic information from participants. These surveys were also implemented on both Apple's iPhone 6s Plus (iOS app version 13.28.06) and Samsung Galaxy Tab 4 (Android app version 1.0.38).

3.3 Statistical Analyses

IBM's Statistical Package for the Social Sciences (SPSS) version 23 for Mac was used to analyse the data. A combination of various statistical procedures were carried out on the data; descriptive analyses, mixed-design analysis of variance (ANOVA) and a two-way analysis of variance (ANOVA) to further explore interactions. The level of $\alpha < .05$ was used for all statistical tests of this experiment.

3.4 Design

The experiment employed a 2 × 2 × 6 mixed-design analysis of variance (ANOVA) with mobile device (Apple's iPhone 6s Plus or Samsung's Galaxy Tab 4) as the between-subjects factor, × time of day (Morning or Afternoon) × time on task (1-min; 5-min; 10-min; 15-min; 20 min; or 25-min) as the within-subjects factors. The morning session (i.e., before 11:00) and the afternoon session (i.e., after 17:00) were held on two consecutive days and counterbalanced.

3.5 Procedure

In order to ensure participants were fully aware of the inclusion and exclusion criteria, all participants were contacted using Cardiff University's Experimental Management System (EMS) emailing system 48 h prior to participation and further reminded 24 h before the start, in addition to being provided with brief instructions through EMS. The study was administered using mobile devices. Participants were either assigned to using an iPhone 6s Plus or a Samsung Galaxy Tab 4. To increase validity and standardisation, all instructions were administered to participants in written form for both the morning and the afternoon session. This study consisted of two parts. The first part was the mobile Psychomotor Vigilance Task (m-PVT) reaction time test, which was a modified version of the Dinges and Powell's [33] Psychomotor Vigilance Task. The m-PVT was run on the Dolphin Web Browser mobile application. The second part was the demographic questionnaire that was distributed within Qualtrics Surveys mobile application. In this modified version, the mobile Psychomotor Vigilance Task (m-PVT) (see Fig. 1), participants were presented with on-screen instructions and a button at the end that read 'Start'. In each trial, participants were shown a black screen background, and at the centre of the screen they would be presented with a large red fixation circle. The red fixation circle (i.e., inter-stimulus interval) would remain on the screen for a randomised duration that lasted between 2–10 s, which was then followed by a yellow stimulus counter. As soon as the inter-stimulus interval reached the randomised duration, a yellow stimulus counter appeared counting up in milliseconds from 0–5 s where it would lapse (i.e., error of omission for 0.5 s) and begin the next trial, or until the participant tapped on the screen. Once the participant tapped on the screen, their reaction time (i.e., stimulus) would be displayed for 0.5 s. At the end of each trial, a black background would appear on-screen for 0.5 s. There were 205 trials in total that lasted approximately 25 min. Kribbs and Dinges [54] found that after a maximum of three trials, the practice effect for the PVT was removed. This study conservatively implemented five practice trials to ensure participants were fully aware of the task, which were removed from final analyses. If participants responded prematurely during any trial (i.e., before the timer commenced counting up), the trial would reset. To also ensure participants were made aware of their premature response, the following message in red was displayed on the centre of the screen, 'You clicked too early! This trial will be reset.' A visual illustration of the mobile Psychomotor Vigilance Task (m-PVT) is presented in Fig. 2.

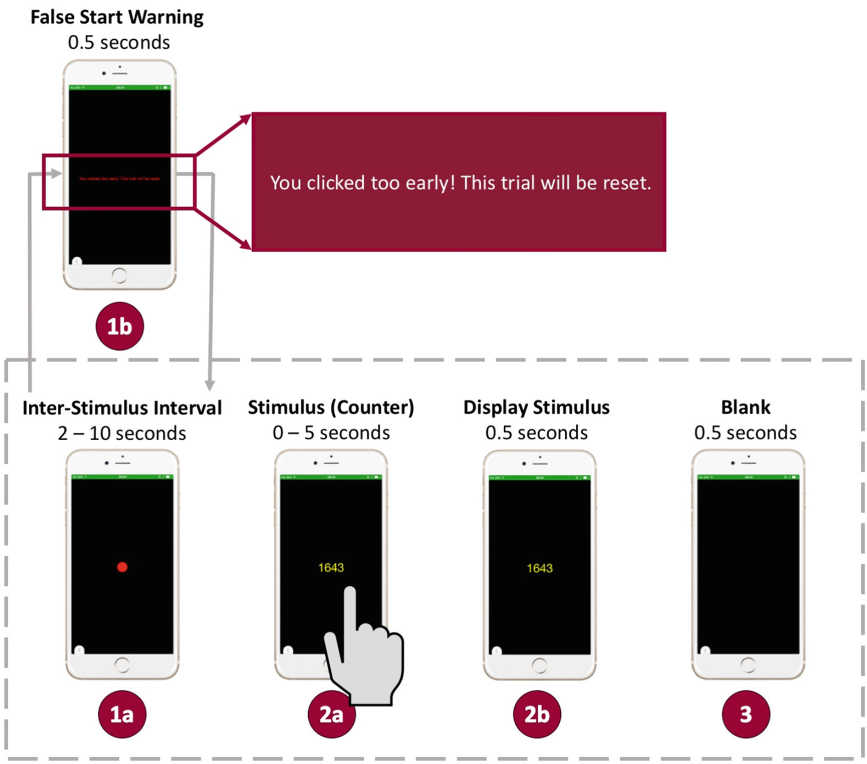


Fig. 1. Mobile Psychomotor Vigilance Task (m-PVT) timeline.

1a. Participants were presented with a large red circle (i.e., inter-stimulus interval), which appeared for a randomised duration between 2–10 s.
1b. If participants responded prematurely, a false start warning message appeared informing them that they clicked too early and that the trial would be reset.
2a. As soon as the inter-stimulus interval reached the randomised duration, a yellow stimulus counter appeared counting up in milliseconds from 0–5 s where it would lapse (i.e., error of omission for 0.5 s) and begin the next trial, or until the participant had tapped on the screen.
2b. Once the participants had tapped on the screen, their reaction time (i.e., stimulus) would be displayed for 0.5 s.
3. At the end of each trial, a black background would appear on-screen for 0.5 s.

Fig. 2. Visual illustration of the mobile Psychomotor Vigilance Task (m-PVT)

4 Results and Discussion

The aim of the study was to investigate whether an alternative mobile version of the 'gold standard' Psychomotor Vigilance Task (PVT) could be used to provide an objective indicator of fatigue in staff working in applied safety critical settings such as train driving, hospital staffs, emergency services, law enforcerments, etc. IBM's Statistical Package for the Social Sciences (SPSS) version 23 for Mac was used to analyse the data. A total of 10,452 test trials were submitted for data analyses, with all 260 practice trials (i.e., 5 practice trials per session) excluded from final analyses. It is important to note that all mobile devices running the online mobile version of the Psychomotor Vigilance Task (m-PVT) were administered through the Dolphin internet browser and were connected using Cardiff University's Eduroam Wi-Fi roaming service. Therefore, on rare occasions when the Wi-Fi connectivity dropped, the participant's trial was lost and thus not recorded. As a result, a total of 1.95% (n = 208) test trials of all 10,660 trials (i.e., 260 practice and 10,400 test) were lost and not recorded. Based on Basner and Dinges [40] recommendations, all 10,452 test trials with reaction time (RTs) < 100 ms (i.e., false start), which accounted for .05% (n = 5) and RTs 500 ms (i.e., number of lapses), which accounted for 31.84% (n = 3,328), were considered for exclusion from the final mean speed response (1/RT) and mean reaction time (RT) analyses. All 31.84% (n = 3,328) of RTs $\geq$ 500 ms (i.e., number of lapses) were analysed separately.

4.1 Mean Speed Response (1/RT) and Reaction Time (RT)

Figure 3 presents the illustrated mean speed responses (1/RTs) across the different conditions while Fig. 4 presents the illustrated mean reaction times (RTs) across the different conditions. Both the 1/RTs and RTs were submitted to a 2 × 2 × 6 mixed-design analysis of variance (ANOVA) with 2 × mobile devices (iPhone 6s Plus or Samsung Galaxy Tab 4) as the between-subjects factor, and × 2 time of day (Morning, or Afternoon) × 6 time on task (1-min; 5-min, 10-min, 15-min, 20 min, or 25-min) as the within-subjects factors. Both the 1/RTs and RTs were significant when comparing the main effect of the two groups using different mobile devices, $F(1, 24)$, 87.21, $p < .001$, $\eta_p^2 = .78$, indicating a large effect size [55, 56] and $F(1, 24)$, 131.85, $p < .001$, $\eta_p^2 = .85$, also indicating a large effect size [55, 56], respectively. In addition, there was a significant main effect of time on task for both the 1/RTs and RTs, Wilks' Lambda = .22, $F(5, 20)$, 14.08, $p < .001$, $\eta_p^2 = .78$, indicating a large effect size [55, 56] and Wilks' Lambda = .24, $F(5, 20)$, 12.66, $p < .001$, $\eta_p^2 = .76$, indicating a large effect size [55, 56], respectively. Furthermore, there was also a significant interaction between mobile devices × time on task for both the 1/RTs and RTs, Wilks' Lambda = .34, $F(5, 20)$, 7.95, $p < .001$, $\eta_p^2 = .67$, indicating a large effect size [55, 56] and Wilks' Lambda = .43, $F(5, 20)$, 5.23, $p = .003$, $\eta_p^2 = .57$, indicating a moderate effect size [55, 56], respectively. The other main effect (time of day) and interactions (two-way interaction, time of day × time on task; and three-way interaction, mobile devices × time of day × time on task) for both 1/RTs and RTs were not significant.

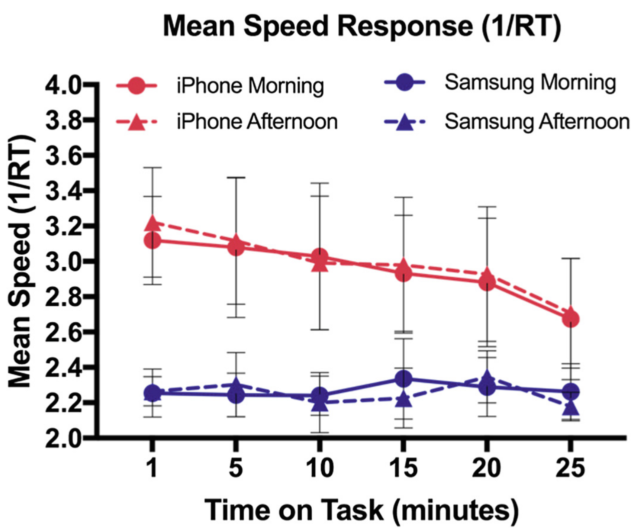


Fig. 3. Mean speed responses (1/RTs) across the different conditions (i.e., morning and afternoon) for both the iPhone 6s Plus and the Samsung Galaxy Tab 4 of the mobile Psychomotor Vigilance Task (m-PVT). *Note:* Mean 1/RTs for both the iPhone 6s Plus and the Samsung Galaxy Tab 4 are presented in bins of 5 min as well as the first minute. Error bars represents standard deviation.

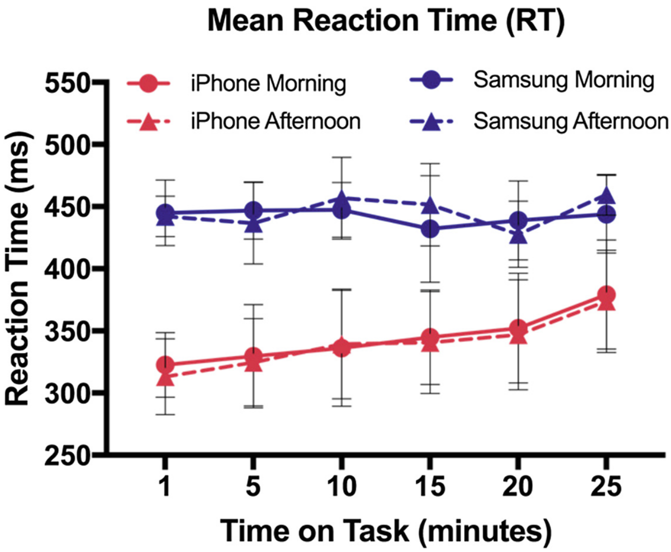


Fig. 4. Mean reaction times (RTs) across the different conditions (i.e., morning and afternoon) for both the iPhone 6s Plus and the Samsung Galaxy Tab 4 of the mobile Psychomotor Vigilance Task (m-PVT). *Note:* Mean RTs for both the iPhone 6s Plus and the Samsung Galaxy Tab 4 are presented in bins of 5 min as well as the first minute. Error bars represents standard deviation.

The main effect of the two groups using different mobile devices was followed by post-hoc tests with Bonferroni correction for multiple comparisons. Post-hoc tests showed that participants' mean speed responses (1/RTs) were significantly greater with the iPhone 6s Plus mobile device ($M = 2.97$, $SE = .05$) than the Samsung Galaxy Tab 4 mobile device ($M = 2.26$, $SE = .05$, $p < .001$). In addition, post-hoc tests also showed that participants' reaction times (RTs) were significantly faster with the iPhone 6s Plus mobile device ($M = 341.92$ ms, $SE = 6.29$ ms) than the Samsung Galaxy Tab 4 mobile device ($M = 444.02$ ms, $SE = 6.29$ ms, $p < .001$). These findings seem to indicate that the iPhone 6s Plus generated significantly greater mean speed responses (1/RTs) and significantly faster mean reaction times (RTs) than the Samsung Galaxy Tab 4, with a mean RT difference of 102 ms between the iPhone 6s Plus and the Samsung Galaxy Tab 4. Therefore, under these circumstances, the interaction between mobile devices × time on task was explored separately with a two-way repeated analysis of variance (ANOVA).

iPhone 6s Plus Mean Speed Response (1/RT) and Reaction Time (RT)

Figures 5 and 6 present the illustrated mean speed of responses (1/RTs) and mean reaction times (RTs) for the iPhone 6s Plus mobile Psychomotor Vigilance Task (m-PVT) across the different conditions. Both the 1/RTs and RTs were submitted to a 2 × 6 two-way repeated analysis of variance (ANOVA) comparing 2 × time of day (Morning, or Afternoon) × 6 time on task (1-min; 5-min, 10-min, 15-min, 20 min, or 25-min). Only the main effect of time on task was significant for both the 1/RTs and RTs, Wilks' Lambda = .12, $F(5, 8)$, 12.02, $p = .001$, $\eta_p^2 = .88$, indicating a large effect size [55, 56] and Wilks' Lambda = .12, $F(5, 8)$, 11.93, p = .002, $\eta_p^2 = .88$, indicating a large effect size [55, 56], respectively. The other main effect (time of day) and interactions (two-way interaction, time of day × time on task) for both 1/RTs and RTs were not significant.

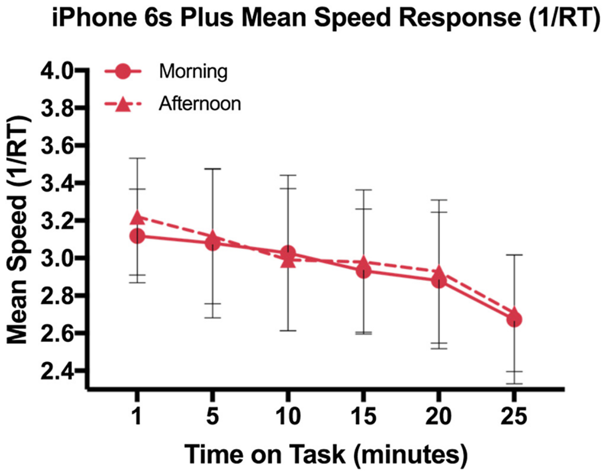


Fig. 5. Mean speed responses (1/RTs) of both the morning session and afternoon session for the iPhone 6s Plus mobile Psychomotor Vigilance Task (m-PVT). *Note:* Mean 1/RTs of both the morning session and afternoon session for the iPhone 6s Plus are presented in bins of 5 min as well as the first minute. Error bars represents standard deviation.

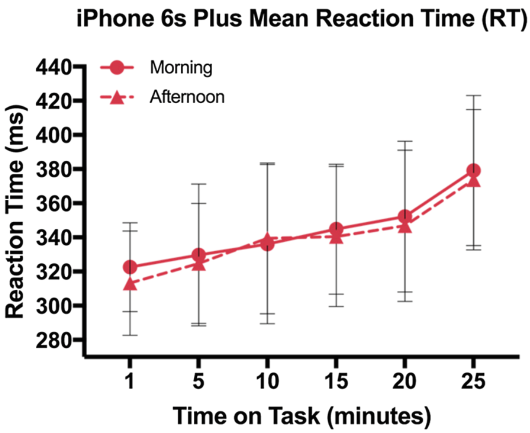


Fig. 6. Mean reaction times (RTs) of both the morning session and afternoon session for the iPhone 6s Plus mobile Psychomotor Vigilance Task (m-PVT). *Note:* Mean RTs of both the morning session and afternoon session are presented in bins of 5 min as well as the first minute. Error bars represents standard deviation.

The main effect of time on task was further explored using Fisher's Least Significant Difference (LSD) post-hoc multiple pairwise comparison, which according to Rovai, Baker and Ponton [57] is used when sample sizes are small. As can be seen from Fig. 7, participants who were assigned to the iPhone mobile device group had significantly greater mean speed responses (1/RTs) between the first minute on the m-PVT (M = 3.17, SE = .07) and 15-min on the m-PVT (M = 2.96, SE = .09, p = .005). In addition, participants had significantly greater 1/RTs between the first minute (M = 3.17, SE = .07) and 20-min (M = 2.90, SE = .10, p = .005). Furthermore, participants had significantly greater 1/RTs between the first minute (M = 3.17, SE = .07) and 25-min (M = 2.69, SE =.07, p < .001). Fisher's LSD post-hoc multiple pairwise comparison also showed potential differences between the first minute on the m-PVT (M = 3.17, SE = .07) and 10-min on the m-PVT (M = 3.01, SE = .10, p = .051). However, this was not statistically significant with this study size. As can be seen from Fig. 8, participants had significantly faster mean reaction times (RTs) between the first minute on the m-PVT (M = 317.89 ms, SE = 7.09 ms) and 10-min on the m-PVT (M = 337.75 ms, SE = 10.27 ms, p = .032). In addition, participants had significantly faster RTs between the first minute (M = 317.89 ms, SE = 7.09 ms) and 15-min (M = 342.70 ms, SE = 10.22 ms, p = .003). Furthermore, participants had significantly faster RTs between the first minute (M = 317.89 ms, SE = 7.09 ms) and 20-min (M = 349.52 ms, SE = 11.42 ms, p = .005). Moreover, participants had significantly faster RTs between the first minute (M = 317.89 ms, SE = 7.09 ms) and 25-min (M = 376.47 ms, SE = 9.20 ms, p < .001).

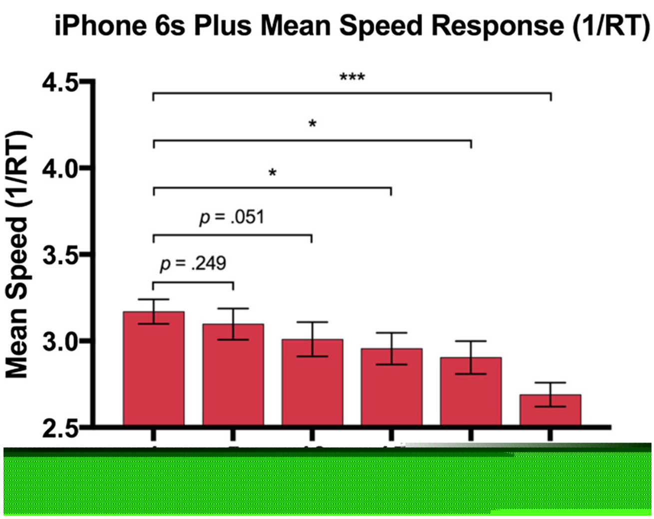


Fig. 7. *p < .05; **p < .005; ***p < .001. *Note:* Mean speed responses (1/RTs) for the iPhone 6s Plus are presented in bins of 5 min as well as the first minute. Error bars represents standard errors.

Teil 1 - Beanspruchungsstruktur

Im Folgenden werden jeweils zwei der sechs Beanspruchungsdimensionen in verschiedenen Kombinationen gegenübergestellt. Geben Sie jeweils an, welche Beanspruchungsdimension für die Gesamtbeanspruchung, die Sie empfunden haben, bedeutsamer war. Es geht also zunächst nicht darum, wie hoch Sie die Beanspruchung in der einzelnen Dimension empfanden, sondern wie wichtig die jeweilige Dimension für das Gesamtempfinden war.

Beispiel:
Wenn für Sie die geistigen Anforderungen, die die Aufgabe gestellt hat, bedeutsamer für das Beanspruchungserleben waren, als die zeitlichen Anforderungen, die Sie empfanden, kreuzen Sie bitte so an:

Geistige Anforderungen	☒	☐	Zeitliche Anforderungen

Körperliche Anforderungen	☐	☐	Zeitliche Anforderungen
Anstrengung	☐	☐	Geistige Anforderungen
Frustration	☐	☐	Körperliche Anforderungen
Anstrengung	☐	☐	Frustration
Geistige Anforderungen	☐	☐	Zeitliche Anforderungen
Körperliche Anforderungen	☐	☐	Anstrengung
Zeitliche Anforderungen	☐	☐	Ausführung der Ausgaben
Frustration	☐	☐	Geistige Anforderungen
Zeitliche Anforderungen	☐	☐	Frustration
Ausführung der Aufgaben	☐	☐	Anstrengung
Anstrengung	☐	☐	Zeitliche Anforderungen
Frustration	☐	☐	Ausführung der Aufgaben
Ausführung der Aufgaben	☐	☐	Körperliche Anforderungen
Geistige Anforderungen	☐	☐	Ausführung der Aufgaben
Geistige Anforderungen	☐	☐	Körperliche Anforderungen

Kontrollieren Sie bitte, ob Sie kein Vergleichspaar vergessen haben.

Bundesanstalt für Arbeitsschutz und Arbeitsmedizin FB 3, Gruppe 3.4

Teil 2 - Beanspruchungshöhe

Geben Sie jetzt bitte an, wie hoch die Beanspruchung in den einzelnen Dimensionen bezogen auf die gerade gelöste Aufgabe ausgeprägt war. Markieren Sie dazu auf den folgenden Skalen bitte, in welchem Maße Sie sich in den sechs genannten Dimensionen von der Aufgabe beansprucht oder gefordert gesehen haben. Machen Sie dafür, wie im Beispiel dargestellt, Ihre Einschätzung mit einem Kreuz kenntlich. Bei Unklarheiten wenden Sie sich bitte an den anwesenden Versuchsleiter.

Beispiel:

x

gering hoch

Geistige Anforderungen

gering hoch

Körperliche Anforderungen

gering hoch

Zeitliche Anforderungen

gering hoch

Ausführung der Aufgaben

gut schlecht

Anstrengung

gering hoch

Frustration

gering hoch

ID:......................... | Datum: .... | VL: .... Seite **4** von **4**

B ISA Questionnaire

Bundesanstalt für Arbeitsschutz und Arbeitsmedizin FB 3, Gruppe 3.4

ID:________________

ISA zur Erfassung der Arbeitsbeanspruchung

Geben Sie jetzt bitte an, wie hoch Ihre Beanspruchung in den letzten 5 Minuten war. Machen Sie dafür Ihre Einschätzung mit einem Kreuz kenntlich. Bei Unklarheiten wenden Sie sich bitte an den anwesenden Versuchsleiter.

				
Der Lotse hat sehr wenig bis gar nichts zu tun.	Der Lotse hat mehr als erforderlich, um die Aufgaben zu erfüllen. Die Zeit vergeht langsam.	Der Lotse hat ausreichend Arbeit. Alle Aufgaben sind unter Kontrolle.	Der Lotse ist „am Limit“. Bestimmte nicht-zwingend notwendige Aufgaben werden verschoben. Die Zeit vergeht schnell	Der Lotse ist überlastet. Einige Aufgaben können nicht erledigt werden. Der Lotse spürt, dass er nicht die Kontrolle hat

ID:........................ | Datum: .... | VL: .... Seite **1** von **2**

References

1. Averty, P., Collet, C., Dittmar, A., Athènes, S., Vernet-Maury, E.: Mental workload in air traffic control: an index constructed from field tests. Aviat. Space Environ. Med. **75**(4), 333–341 (2004)
2. Beatty, J.: Task-evoked pupillary responses, processing load, and the structure of processing resources. Psychol. Bull. **91**(2), 276–292 (1982)
3. Berster, P., et al.: Luftverkehrsbericht 2016 - Daten und Kommentierungen des deutschen und weltweiten Luftverkehrs (2017). http://www.dlr.de/fw/Portaldata/42/Resources/dokumente/pdf/__Luftverkehrsbericht_2016_final_141217.pdf
4. Brennan, S.: An experimental report on rating scale descriptior sets for the instantaneous self assessment (ISA) recorder. Technical report. DRA Technical Memorandum (CAD5) 92017, DRA Maritime Command and Control Division (1992)
5. Cain, B.: A review of the mental workload literature. Technical report, Defence Research and Development Canada, Human System Integration Section, Toronto (2007). http://ftp.rta.nato.int/public/pubfulltext/rto/tr/rto-tr-hfm-121-part-ii/tr-hfm-121-part-ii-04.pdf. Accessed 11 Dec 2013
6. Edwards, T., Martin, L., Bienert, N., Mercer, J.: The relationship between workload and performance in air traffic control: exploring the influence of levels of automation and variation in task demand. In: Longo, L., Leva, M.C. (eds.) H-WORKLOAD 2017. CCIS, vol. 726, pp. 120–139. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61061-0_8
7. Eggemeier, F., Wilson, G.F., Kramer, A.F., Damos, D.L.: Workload assessment in multi-task environments. In: Multiple-task Performance, pp. 207–216. Taylor & Francis (1991)

8. Hancock, P.A.: Whither workload? Mapping a path for its future development. In: Longo, L., Leva, M.C. (eds.) H-WORKLOAD 2017. CCIS, vol. 726, pp. 3–17. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61061-0_1
9. Hancock, P.A., Desmond, P.A.: Stress, Workload, and Fatigue. Lawrence Erlbaum Associates Publishers, Hillsdale (2001)
10. Hart, S.G., Staveland, L.E.: Development of the NASA TLX: results of empirical and theoretical research. In: Hancock, P., Meshkati, N. (eds.) Human Mental Workload, pp. 139–183, North Holland, Amsterdam (1988)
11. Hendy, K.C., Hamilton, K.M., Landry, L.N.: Measuring subjective workload: when is one scale better than many? Hum. Factors **35**, 579–601 (1993)
12. Hilburn, B., Jorna, P.G.A.M.: Workload and air traffic control. In: Hancock, A., Desmond, P. (eds.) Stress, Workload, and Fatigue, pp. 384–394. Erlbaum, Mahwah (2001)
13. Jordan, C.: Experimental study of the effect of an instantaneous self assessment workload recorder on task performance. Technical report. DRA Technical Memorandum (CAD5) 92011, DRA Maritime Command Control Division (1992)
14. Kahneman, D.: Attention and Effort. Prentice-Hall, Englewood Cliffs (1973)
15. Kirwan, B., et al.: Human factors in the ATM system design life cycle. In: FAA/Eurocontrol ATM R&D Seminar (1997)
16. Matthews, G.: Levels of transaction: a cognitive sciences framework for operator stress, pp. 5–33. Lawrence Erlbaum Associates Publishers Mahwah (2001)
17. Meister, D.: Behavioral Foundations of System Development. Wiley, New York (1976)
18. Mulder, L.J.M., van Roon, A.M., Althaus, M., Laumann, K., Dicke, M.: Determining dynamic cardiovascular state changes using a baro-reflex simulation model. In: Human factors in transportation, communication, health and the workplace, pp. 297–315. Shaker Publishing, Maastricht (2002)
19. O'Donnell, C.R.D., Eggemeier, F.T.: Workload assessment methodology. In: Boff, K., Kaufmann, L., Thomas, J. (eds.) Handbook of Perception and Human Performance, pp. 42–49. Wiley (1989). http://apps.usd.edu/coglab/schieber/docs/odonnell.pdf, Accessed 19 June 2015
20. Parasuraman, R., Hancock, P.: Adaptive control of workload. In: Hancock, A., Desmond, P. (eds.) Stress, Workload, and Fatigue, pp. 305–320. Erlbaum, Mahwah (2001)
21. Pribram, K.H., McGuinness, D.: Arousal, activation, and effort in the control of attention. Psychol. Rev. **82**, 116–149 (1975). http://www.karlpribram.com/wp-content/uploads/pdf/theory/T-068.pdf. Accessed on 15 May 2015
22. Radüntz, T.: Dual frequency head maps: a new method for indexing mental workload continuously during execution of cognitive tasks. Front. Physiol. **8**, 1019 (2017). https://www.frontiersin.org/article/10.3389/fphys.2017.01019
23. Saxby, D.J., Matthews, G., Warm, J.S., Hitchcock, E.M., Neubauer, C.: Active and passive fatigue in simulated driving: discriminating styles of workload regulation and their safety impacts. J. Exp. Psychol.: Appl. **19**(4), 287–300 (2013)
24. Sonnentag, S., Kruel, U.: Psychological detachment from work during off-job time: the role of job stressors, job involvement, and recovery-related self-efficacy. Eur. J. Work. Organ. Psychol. **15**(2), 197–217 (2006). https://doi.org/10.1080/13594320500513939
25. Stanton, N., Salmon, P., Walker, G., Baber, C., Jenkins, D.: Human Factors Methods: A Practical Guide for Engineering and Design, December 2005

26. Ullsperger, P., Metz, A.M., Gille, H.G.: The P300 component of the event-related brain potential and mental effort. Ergonomics **31**(8), 1127–1137 (1988). https://doi.org/10.1080/00140138808966752. pMID: 3191898
27. Veltman, J.A., Gaillard, A.W.K.: Physiological indices of workload in a simulated flight task. Biol. Psychol. **42**(3), 323–342 (1996). http://www.sciencedirect.com/science/article/pii/0301051195051651
28. Vidulich, A.M., Tsang, P.: Mental workload and situation awareness, pp. 243–273, 4th edn. Wiley, Hoboken (2012)
29. de Waard, D.: The measurement of drivers' mental workload. Ph.D. thesis, University of Groningen, Traffic Research Centre, Haren, Netherlands (1996)
30. Wickens, C.D.: Multiple resources and performance prediction. Theor. Issues Ergon. Sci. **3**(2), 159–177 (2002). http://www.tandfonline.com/doi/abs/10.1080/14639220210123806
31. Wickens, C.D.: Mental workload: assessment, prediction and consequences. In: Longo, L., Leva, M.C. (eds.) H-WORKLOAD 2017. CCIS, vol. 726, pp. 18–29. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61061-0_2
32. Xie, B., Salvendy, G.: Review and reappraisal of modelling and predicting mental workload in single- and multi-task environments. Work Stress **14**(1), 74–99 (2000)
33. Yerkes, R.M., Dodson, J.D.: The relation of strength of stimulus to rapidity of habit-formation. J. Comp. Neurol. Psychol. **18**, 459–482 (1908). http://psychclassics.yorku.ca/Yerkes/Law. Accessed 03 Nov 2011

The Effect of Education and Training on Mental Workload in Medical Education

Aidan Byrne[1,2,3](✉)

[1] ABM Local Health Board, Swansea, UK
aidanbyrne27@swansea.ac.uk
[2] School of Medicine, Swansea University, Swansea, UK
[3] Department of Anaesthesia, Morriston Hospital, Swansea SA4 4NL, UK

Abstract. Mental workload is now accepted as a significant factor in the performance of individuals conducting complex tasks in high risk environments such as healthcare and research shows that experience reduces the levels of mental workload. In order to determine the effect of training on mental workload, a systematic search of studies of the effect of education and/or training was conducted using standard research databases. Only 6 studies were identified and these showed either limited or no improvement after training/education, suggesting that educational interventions have limited utility in reducing mental workload. The apparent failure of education may suggest that the current, digital view of cognition is not adequate and an alternative concept of analogue cognition is proposed as a possible explanation.

Keywords: Mental workload · Education · Training · Human performance

1 Introduction

Mental workload is an abstract concept which reflects the demands of a particular task/environment on the cognitive resources of individuals [1, 2]. One of the main beliefs underpinning the study of mental workload as a concept is that high levels of mental workload lead to poor performance [3] and that interventions such as training/education can reduce mental workload and therefore improve performance [4]. In addition, measurement of mental workload could be used to develop improved educational techniques [5, 6]. Although education can be seen as peripheral to the more scientific process of characterising and measuring mental workload, the concept of mental workload would remain an abstract if it cannot be applied to real world applications [7, 8]. Although this form or research can be used to modify any specific task or the environment in which it is performed, it is a change in human behaviour that provides the greatest challenge to promoting safety [9, 10].

Within medicine, the concept of mental workload is not yet widely used to aid the design of either instructional materials or assessment tools, [6] although it is now being used to aid the design of systems and equipment [11, 12]. In addition, the concept of cognitive load theory has been introduced which, as the theories share so much, can lead to confusion [13, 14]. While Cognitive Load theory has mainly analysed classroom based teaching and is based on the premise that a reduction in the cognitive load

L. Longo and M. C. Leva (Eds.): H-WORKLOAD 2018, CCIS 1012, pp. 258–266, 2019.
https://doi.org/10.1007/978-3-030-14273-5_15

while learning provides learners with an increased cognitive capacity to learn. In contrast, mental workload theory has mainly analysed operators working in real or realistic environments and is based on the premise that either reducing the workload associated with the task or increased training will allow subjects to improve their performance. In addition, while mental workload theory posits mental workload to be a rather complex, multidimensional concept, cognitive load theory has the additional complication of splitting cognitive load into three separate components, making measurement difficult. This paper deals with mental workload theory rather than cognitive load theory.

Studies which measured the difference between novices and experts, completing tasks in simulated or real medical environments have shown that experience does reduce cognitive load [15–18]. However, studies within medical education which have measured the workload of subjects before and after a specific episode of training are problematic in that it appears to suggest that training does not reliably reduce workload [19–21].

If we accept the current definition of mental workload as the balance between the cognitive demands of the task compared to the ability of the subject [1] and that the purpose of training is to improve ability [6], then it follows that training should reduce mental workload. However, if training medical staff does not reduce their mental workload then either the training is ineffective or the tasks undertaken by medical practitioners have not been adequately characterised. This question is important as while mental workload is now accepted as a valid outcome in the assessment of both individual performance and that of systems in healthcare, [3, 22, 23] workers are currently in short supply and in many cases highly paid, so the allocation of staff to training rather than patient care provides a significant increase in cost or a reduction in patient care. In such circumstances, it is essential that any educational interventions are demonstrably effective. An initial search of the medical educational literature did not provide adequate evidence to address the question of whether training affects the mental workload of operators, so a wider search of the literature was undertaken. The aim was to identify all published papers which measured the effect of education/training on mental workload to determine which forms of training are most effective in reducing mental workload.

2 Methods

A systematic search of online research databases was performed by a single author, using the terms “mental workload” and “training” or “education”. Inclusion criteria were any study which included an objective measure of mental workload before and after a period of training or education. Any form of educational intervention was included. Studies were excluded if only subjective measures of mental workload were used. The number of studies identified by each search engine are shown in Table 1. The abstract for each study identified was read to determine whether it was suitable for inclusion within this paper, with total numbers included also shown in Table 1. On full review of all the papers, none needed to be excluded. After the initial search, any authors identified were then used as search items to determine if they has authored similar papers, but this did not identify any additional studies. A search using Google Scholar using the terms "mental workload" and "training" produced 23,800 results. An informal search of the first 500 results did not reveal any further studies which met the criteria. Papers were then analysed to determine the task studied, the type and duration of educational intervention and the effect of mental workload.

3 Results

The search process identified six studies which met the inclusion criteria (Tables 1 and 2). A large number of papers were common to many of the database results, supporting the conclusion that the search identified all studies with the specified items in the title or abstract. However, it is likely that there are other studies which measure mental workload before and after some educational intervention, but which have not been identified by this process. The following six studies were identified:

1 - Lelis-Torres et al. [24] studied mental workload using a processed EEG signal before and after subjects practiced a task which involved repetitive key pressing and showed that the subject's workload decreased. The task or training were not well described but appear to be highly constrained with simple practice rather than a designed educational intervention. They found that repetitive practice resulted both in lower initial workload and a greater fall in workload with practice. However, they conclude that the increased cognitive effort involved in random practice showed more potential as an educational intervention.

2 - Boet et al. [21] performed a randomised trial of training in 20 junior surgical residents being trained on a full scale simulator to deal with medical crises. Mental workload was measured using a vibrotactile secondary task method. In the control group (n = 10) practicing crisis management without debriefing appeared to increase workload. In the intervention group (n = 10) targeted feedback over 15 min which included teaching on generic crisis management skills reduced workload.

3 - Byrne et al. [19] studied the performance of doctors being trained to act as examiners in a checklist based practical skills examination. Mental workload was measured by a vibrotactile secondary task method. Subjects completed four sequential video based training sessions with feedback. Overall, training appeared to have no effect on mental workload or the accuracy of their marking.

4 - Byrne et al. [25] studied anaesthetic trainees being trained on a full scale training simulator. Workload was measured with a vibrotactile secondary task method. After a single simulation with debriefing and targeted advice, each subject then completed a further simulation. There was no effect on mental workload or performance.

5 - Saus et al. [26] studied police cadets being trained to use firearms in a simulated environment. The training focussed on developing situational awareness on the

use of firearms in a variety of simulated situations. Mental workload was measured by a reduction in heart rate variability with the signal generated by a surface ECG. A single training session reduced mental workload and improved performance measured by the number of shots fired and the number of hits.

6 - Kang et al. [27] studied college students to complete simple mathematical calculations similar to "If A = 3, 7 + A = ?". Mental workload was measured by thermography of the nose. Although the measure appeared to correlate with perceived workload, there was no improvement during the short practice phase.

4 Discussion

Despite the large volume of literature on mental workload, there are few studies which have measured the effect of training or education on mental workload measured by an objective methodology. Including subjective measures [28] would have considerably increased the number of studies included in this paper, but as training/education would be expected to affect subjective measures of workload they were not included. The methodologies used to measure workload were diverse and although each methodology was supported by appropriate references to validate its use, the evidence for each appeared to be limited. In addition, it is recognised that to be effective, educational interventions need to use appropriate instructional design techniques, [29] usually incorporating repeated practice and feedback [30]. In contrast, the studies identified here used single educational interventions which were both short and appeared to lack any educational theory or design. They also appeared to focus on unaided practice or the provision of simple instruction.

Overall, the effect of training/education on mental workload appears to be variable, so that a logical conclusion from this review might be that there was insufficient evidence to recommend training/education as a valid intervention. However, two of the tasks studied were very limited, repetitive tasks, one of which showed improvement. This suggests that for such simple tasks, repetition may be effective. The other four studies looked at complex, rapidly evolving environments with the training either targeted at dealing with the specific scenario, or with training targeting generic crisis management or situational awareness. In these more complex situations, training staff to deal with specific problems had no effect, but where generic skills were emphasised, workload decreased.

Where the mental workload of medical staff, who work in complex environments, at different stages of their careers have been measured, it appears to show a gradual decrease in workload over many years of training [15, 17]. This gradual decrease in mental workload over many years of practice is not compatible with many existing models of learning, such as competency based training, where students acquire specific knowledge and skills during defined learning periods, with mastery of each component tested at the end of each section with a summative assessment. It also raises questions as to how this learning should be represented in models of neural processing.

Most educational theory posits cognition, at least implicitly, as a digital process using discreet items of knowledge acquired from sensory input which is then processed via a logical, Boolean or Bayesian logic process. The inferences from this conceptualisation of learning is that if a subject has acquired adequate knowledge relating to a specific situation, then they will be likely to select the correct solution to any given problem. Correspondingly, if the subject makes an incorrect choice, then they must have inadequate knowledge which needs to be addressed through an educational process.

An alternative view is to consider human cognition to be based on massive numbers of analogue circuits which integrate sensory inputs and develop responses in the same way as the electronic analogue computers used in the 1950/60s [31]. Although the term 'analogue' when applied to a computer describes the relationship between input and output of the computation, it does not specify the internal construction of the device, which can use, for example, physical connections or electrical circuits. Clearly, the human brain does not contain either physical linkages or continuous electrical currents. However, while neurones cannot conduct continuous currents which are turned up or down, they can form circuits which increase their firing rate or 'resonate' in response to specific conditions. The description of perception in terms of circuits which 'resonate' or 'vibrate' in response to incoming sensory information has already been described in Cognitive Resonance Theory [32]. This analogue view of learning would explain the learning process as one of gradual 'tuning' of circuits to improve their 'signal to noise ratio' and 'quality of reception' more in keeping with gradual improvements to an AM radio or the gradual redesign of a violin so that each becomes more responsive to any input and produces a progressively higher quality output [33].

Such a view of cognition would explain the lack of evidence to demonstrate that education is effective in reducing workload, as it might be expected that performance of simple, highly constrained tasks could be improved after short practice, extensive experience/practice would be required to 'retune' or develop the required analogue circuits where a more complex task required active selection and processing of information from a wide range of incoming sensory input [34]. It would also explain why short training interventions to deal with specific problems might fail, interventions which assisted subjects in making sense of the complexity around them might provide immediate reductions in mental workload, as rather than trying to analyse complexity, subjects could more easily select from a limited number of options. Further, new concepts of education may support the development of new and more effective teaching methods [35, 36].

5 Conclusion

The published literature on the effect of education on the mental workload of operators using objective methodology is extremely limited, with published studies limited to single educational interventions of uncertain utility. In addition, while the performance of highly constrained, simple tasks may be improved by short term practice or debriefing, they appear ineffective when applied to more complex tasks. In addition, while training subjects how to deal with specific complex tasks appears to be ineffective, teaching subjects a generic approach to a range of problems appears more effective.

These results may be better explained by an analogue theory of cognition rather than a digital one. Further research looking at the effects of education on mental workload will be required to determine whether this is true. In particular, studies which involve the measurement of mental workload in realistic environments before, during and after prolonged periods of training/education will be required to determine which forms of education are most effective in reducing mental workload.

Appendix

Table 1. Summary of databases searched and suitable studies identified. Terms used were "mental workload" and "training" or "education"

	Database	Match	Matches	Additional papers included	Total included in analysis
1	Pubmed	Training	153	5	5
2	Pubmed	Education	109	0	5
3	Science direct	Training	64	0	5
4	Science direct	Education	16	0	5
5	CINALH	Training	3	0	5
6	CINALH	Education	1	0	5
7	ERIC	Training	1	0	5
8	ERIC	Education	10	0	5
9	PsychInfo	Training	13	0	5
10	PsychInfo	Education	2	0	5
11	Scopus	Training	192	1	6
12	Scopus	Education	39	0	6
13	Web of science	Training	16	0	6
14	Web of science	Education	2	0	6
15	Wiley interscience	Training	9	0	6
16	Wiley interscience	Education	16	0	6

Table 2. Details of studies included in analysis.

	Study	Primary task	Outcome	Intervention	Duration	Retest	Outcome
1	Lelis-Torres et al. [24]	Sequential key pressing	Processed EEG	Repetitive practice	Post practice	Immediate	Improved
2	Boet et al. [21]	Managing simulated medical crisis	Secondary task	Debriefing training on crisis management	One training session	Immediate	Improved
3	Byrne et al. [19]	Marking skills assessment (OSCE)	Secondary task	Video based practice	One training session	Immediate	No effect
4	Byrne et al. [25, 37]	Simulated anaesthetic crisis	Secondary tasks	Video based debriefing on specific crisis	One training session	Immediate	No effect
5	Kang et al. [27]	Shooting simulator	Heart rate variability	Situational awareness training	One training session	Immediate	Improved
6	Kang et al.	Alphanumeric processing	Thermography	Practice	One session	Immediate	No effect

References

1. Wickens, C.D.: Multiple resources and mental workload. Hum. Factors **50**, 449–455 (2008). https://doi.org/10.1518/001872008X288394
2. Longo, L., Leva, M.C.: Human Mental Workload: Models and Applications: First International Symposium, H-WORKLOAD 2017, Dublin, Ireland, June 28-30, 2017, Revised Selected Papers. Springer, Heidelberg (2017). https://doi.org/10.1007/978-3-319-61061-0
3. Yurko, Y.Y., Scerbo, M.W., Prabhu, A.S., Acker, C.E., Stefanidis, D.: Higher mental workload is associated with poorer laparoscopic performance as measured by the NASA-TLX Tool. Simul. Healthc. J. Soc. Simul. Healthc. **5**, 267–271 (2010). https://doi.org/10.1097/SIH.0b013e3181e3f329
4. Carswell, C.M., Clarke, D., Seales, W.B.: Assessing mental workload during laparoscopic surgery. Surg. Innov. **12**, 80–90 (2005). https://doi.org/10.1177/155335060501200112
5. Menekse Dalveren, G.G., Cagiltay, N.E., Ozcelik, E., Maras, H.: Insights from pupil size to mental workload of surgical residents: feasibility of an educational computer-based surgical simulation environment (ECE) considering the hand condition. Surg. Innov. (2018). https://doi.org/10.1177/1553350618800078
6. Byrne, A.: Mental workload as an outcome in medical education. In: Longo, L., Leva, M.C. (eds.) H-WORKLOAD 2017. CCIS, vol. 726, pp. 187–197. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61061-0_12
7. Parasuraman, R., Sheridan, T.B., Wickens, C.D.: Situation awareness, mental workload, and trust in automation: viable, empirically supported cognitive engineering constructs. J. Cogn. Eng. Decis. Mak. **2**, 140–160 (2008). https://doi.org/10.1518/155534308X284417

8. Ayaz, H., Shewokis, P.A., Bunce, S., Izzetoglu, K., Willems, B., Onaral, B.: Optical brain monitoring for operator training and mental workload assessment. Neuroimage **59**, 36–47 (2012). https://doi.org/10.1016/J.NEUROIMAGE.2011.06.023
9. Young, M.S., Brookhuis, K.A., Wickens, C.D., Hancock, P.A.: State of science: mental workload in ergonomics. Ergonomics **58**(1), 1–17 (2015)
10. Wickens, C.D.: Mental workload: assessment, prediction and consequences. In: Longo, L., Leva, M.C. (eds.) H-WORKLOAD 2017. CCIS, vol. 726, pp. 18–29. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61061-0_2
11. Longo, L.: Mental workload in medicine: foundations, applications, open problems, challenges and future perspectives. In: 2016 IEEE 29th International Symposium on Computer-Based Medical Systems (CBMS), pp. 106–111. IEEE (2016)
12. Longo, L.: Designing medical interactive systems via assessment of human mental workload. In: 2015 IEEE 28th International Symposium on Computer-Based Medical Systems, pp. 364–365. IEEE (2015)
13. Sweller, J.: Cognitive load theory. Psychol. Learn. Motiv. **55**, 37–76 (2011). https://doi.org/10.1016/B978-0-12-387691-1.00002-8
14. van Merriënboer, J.J.G., Sweller, J.: Cognitive load theory and complex learning: recent developments and future directions. Educ. Psychol. Rev. **17**, 147–177 (2005). https://doi.org/10.1007/s10648-005-3951-0
15. Zheng, B., Cassera, M.A., Martinec, D.V., Spaun, G.O., Swanstrom, L.L.: Measuring mental workload during the performance of advanced laparoscopic tasks. Surg. Endosc. **24**, 45–50 (2010). https://doi.org/10.1007/s00464-009-0522-3
16. Guru, K.A., et al.: Cognitive skills assessment during robot-assisted surgery: separating the wheat from the chaff. BJU Int. **115**, 166–174 (2015). https://doi.org/10.1111/bju.12657
17. Byrne, A.J.J., Murphy, A., McIntyre, O., Tweed, N.: The relationship between experience and mental workload in anaesthetic practice: an observational study. Anaesthesia **68**, 1266–1272 (2013). https://doi.org/10.1111/anae.12455
18. Byrne, A.J., et al.: Novel method of measuring the mental workload of anaesthetists during clinical practice. Br. J. Anaesth. **105**, 767–771 (2010). https://doi.org/10.1093/bja/aeq240
19. Byrne, A., Soskova, T., Dawkins, J., Coombes, L.: A pilot study of marking accuracy and mental workload as measures of OSCE examiner performance. BMC Med. Educ. **16**, 191 (2016). https://doi.org/10.1186/s12909-016-0708-z
20. Muresan 3rd, C., Lee, T.H., Seagull, J., Park, A.E.: Transfer of training in the development of intracorporeal suturing skill in medical student novices: a prospective randomized trial. Am. J. Surg. **200**, 537–541 (2010). https://doi.org/10.1016/j.amjsurg.2009.12.018
21. Boet, S., Sharma, B., Pigford, A.-A., Hladkowicz, E., Rittenhouse, N., Grantcharov, T.: Debriefing decreases mental workload in surgical crisis: a randomized controlled trial. Surgery **161**, 1215–1220 (2017). https://doi.org/10.1016/j.surg.2016.11.031
22. Sato, H., Miyashita, T., Kawakami, H., Nagamine, Y., Takaki, S., Goto, T.: Influence of mental workload on the performance of anesthesiologists during induction of general anesthesia: a patient simulator study. Biomed. Res. Int. **2016**, 1058750 (2016). https://doi.org/10.1155/2016/1058750
23. Carayon, P., Gürses, A.P.: A human factors engineering conceptual framework of nursing workload and patient safety in intensive care units. Intensive Crit. Care Nurs. **21**, 284–301 (2005). https://doi.org/10.1016/J.ICCN.2004.12.003
24. Lelis-Torres, N., Ugrinowitsch, H., Apolinário-Souza, T., Benda, R.N., Lage, G.M.: Task engagement and mental workload involved in variation and repetition of a motor skill. Sci. Rep. **7**, 14764 (2017). https://doi.org/10.1038/s41598-017-15343-3
25. Byrne, A.J., et al.: Effect of videotape feedback on anaesthetists' performance while managing simulated anaesthetic crises: a multicentre study. Anaesthesia **57**, 176–179 (2002)

26. Saus, E.-R., Johnsen, B.H., Eid, J., Riisem, P.K., Andersen, R., Thayer, J.F.: The effect of brief situational awareness training in a police shooting simulator: an experimental study (2009). https://doi.org/10.1207/s15327876mp1803s_2
27. Kang, J., Babski-Reeves, K.: Detecting mental workload fluctuation during learning of a novel task using thermography. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 52, pp. 1527–1531 (2008). https://doi.org/10.1177/154193120805201947
28. Rubio, S., Diaz, E., Martin, J., Puente, J.M.: Evaluation of subjective mental workload: A comparison of SWAT, NASA-TLX, and workload profile methods. Appl. Psychol. **53**, 61–86 (2004). https://doi.org/10.1111/j.1464-0597.2004.00161.x
29. Sweller, J., Sweller, J.: Instructional design. Aust. Educ. Rev. (1999)
30. Scerbo, M.W., Britt, R.C., Montano, M., Kennedy, R.A., Prytz, E., Stefanidis, D.: Effects of a retention interval and refresher session on intracorporeal suturing and knot tying skill and mental workload. Surgery **161**, 1209–1214 (2017). https://doi.org/10.1016/j.surg.2016.11.011
31. Small, J.S.: The Analogue Alternative: The Electronic Analogue Computer in Britain and the USA, 1930–1975. Routledge, Abingdon (2001)
32. Grossberg, S.: Adaptive Resonance Theory: how a brain learns to consciously attend, learn, and recognize a changing world. Neural Netw. **37**, 1–47 (2013). https://doi.org/10.1016/J.NEUNET.2012.09.017
33. Ogiela, Lidia: Cognitive systems for medical pattern understanding and diagnosis. In: Lovrek, Ignac, Howlett, R.J., Jain, L.C. (eds.) KES 2008. LNCS (LNAI), vol. 5177, pp. 394–400. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-85563-7_51
34. Abela, J.: Adult learning theories and medical education: a review. Malta Med. J. **21**(1), 11–18 (2009)
35. Gobbo, F., Longo, L., Orru, G., O'sullivan, D.: An investigation of the impact of a social constructivist teaching approach, based on trigger questions, through measures of mental workload and efficiency. In: Proceedings of the 10th International Conference on Computer Supported Education, CSEDU 2018, Funchal, Madeira, Portugal, 15–17 March 2018, vol. 2 (2018)
36. Hancock, P.A.: Whither workload? Mapping a path for its future development. In: Longo, Luca, Leva, M.C. (eds.) H-WORKLOAD 2017. CCIS, vol. 726, pp. 3–17. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61061-0_1
37. Smith, M.A., Byrne, A.J.: 'Help! I need somebody': getting timely assistance in clinical practice. Anaesthesia **71**, 755–759 (2016). https://doi.org/10.1111/anae.13497

Author Index