2018

On the Reliability, Validity and Sensitivity of Three Mental Workload Assessment Techniques for the Evaluation of Instructional Designs: A Case Study in a Third-level Course

Luca Longo
Technological University Dublin, luca.longo@tudublin.ie

Recommended Citation

Longo, L. (2018). On the reliability, validity and sensitivity of three mental workload assessment techniques for the evaluation of instructional designs: a case study in a third-level course. *10th International Conference on Computer Supported Education - CSEDU 2018*, 15-17 March, Funchal, Madeira, Portugal. doi:10.5220/0006801801660178

On the reliability, validity and sensitivity of three mental workload assessment techniques for the evaluation of instructional designs: a case study in a third-level course

Luca Longo
School of Computing, College of Sciences and Health, Dublin Institute of Technology, Dublin, Ireland
ADAPT: The global centre of excellence for digital content and media innovation. Dublin, Ireland
**luca.longo@dit.ie*

Keywords: Human Mental Workload; Cognitive Load Theory; Instructional Design;

Abstract: Cognitive Load Theory (CLT) has been conceived for instructional designers eager to create instructional resources that are presented in a way that encourages the activities of the learners and optimise their performance, thus their learning. Although it has been researched for many years, it has been criticised because of its theoretical clarity and its methodological approach. In particular, one fundamental and open problem is the measurement of its cognitive load types and the measurement of the overall cognitive load of learners during learning tasks. This paper is aimed at investigating the reliability, validity and sensitivity of existing mental workload assessment techniques, borrowed from the discipline of Ergonomics, when applied to the field of Education, Teaching and Learning. In details, a primary research involved the application of three subjective mental workload assessment techniques, namely the NASA Task Load Index, the Workload Profile and the Rating Scale Mental Effort, in a typical third-level classroom for the evaluation of two instructional design conditions. The Cognitive Theory of Multimedia Learning and its design principles have been used as the underlying theoretical framework for the design of the two conditions. Evidence strongly suggests that the three selected mental workload measures are highly reliable within Education and their moderate validity is in line with results obtained in Ergonomics.

1 Introduction

Cognitive Load Theory (CLT) (Sweller et al., 1998) has been conceived as a form of guidance for instructional designers eager to create instructional resources that are presented in a way that encourages the activities of the learners and optimise their performance, thus their learning (Chandler and Sweller, 1991). Although CLT has been researched for many years, providing a series of effects and guidelines to create effective instructional designs, it has been criticised because of its theoretical clarity (Schnotz and Kürschner, 2007) and its methodological approach (Gerjets et al., 2009). In particular, one fundamental and open problem is the measurement of the cognitive load of learners during learning tasks (Paas et al., 2003). Within CLT, three types of cognitive load have been conceptualised: intrinsic, extraneous and germane. These are the fundamental building blocks (the assumptions) of the theory itself. The intrinsic load is influenced by the unfamiliarity of the learners or the intrinsic complexity of the learning material under use. The extraneous load is impacted by the way the instructional material is designed, organised and presented. The germane load is affected by the effort devoted for the processing of information, the construction and automation of schemas in the brain of learners. According to the traditional critical rationalism proposed by Popper (2014), CLT cannot be considered a scientific theory because its fundamental assumptions - the cognitive load types - cannot be measured, tested empirically and therefore they are not falsifiable (Gerjets et al., 2009). Because of this, the scientific value of Cognitive Load Theory and all the other theories built upon the notion of cognitive load (Goldman, 1991; Gerjets et al., 2009) still lack empirical validation. Due to the above reasons, the main research challenge in this area concerns the development of reliable and valid measures of the cognitive load types and the development of overall measures of cognitive load that can be applied in the general field of Education and in the specific field of Teaching and Learning. Another domain in which cognitive load is heavily

researched and employed is Ergonomics (Young et al., 2015). Here, the psychological construct of cognitive load, mainly referred to as human Mental Workload (MWL), has a long history with several applications in the aviation (Hart, 2006) and automotive industries (Brookhuis and de Waard, 2010). In these domains, many measurement techniques, both uni-dimensional and multi-dimensional have been developed for MWL assessment (Cain, 2007; Young et al., 2015). Similarly, various criteria for validating these techniques have been proposed during the last 5 decades, indicating the importance of research on MWL (Rubio et al., 2004). Generally speaking, the main reason for assessing MWL, in Ergonomics, is to measure the mental cost of performing a task with the goal of predicting operator and system performance (Cain, 2007). In Education the situation is similar: the main reason for assessing cognitive load is to measure the mental cost of performing a learning task with the goal of predicting the learner's performance and thus learning.

This paper is an attempt to evaluate the reliability, validity and sensitivity of existing measures of overall mental workload, borrowed from the discipline of Ergonomics, for the evaluation of different instructional design conditions. Three mental workload measures have been selected: the multidimensional Nasa Task Load Index (Hart, 2006) and Workload Profile (Tsang and Velazquez, 1996) as well as the unidimensional Rating Scale Mental Effort (Zijlstra, 1993). A primary research study has been shaped including the comparison of two different instructional design conditions in a third-level master module. The first condition includes the delivery of instructional material in a traditional one-way (lecturer to students) employing slides projected to a white-board and verbally presented to learners. The second condition includes the conversion of the instructional material of the first condition into multimedia videos developed by following a set of design principles from the Cognitive Theory of Multimedia Learning (Mayer, 2002). A schematic summary of the gaps in the literature and the solution proposed are depicted in figure 1.

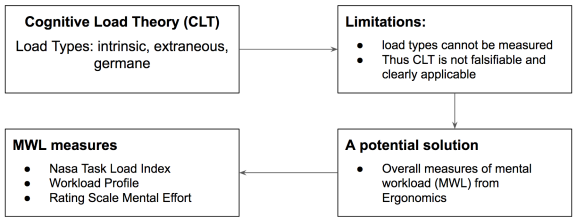

Figure 1: Summary of the research design

The rest of the paper is organised as it follows. Section 2 introduces the theoretical frameworks, including Cognitive Load Theory and its load types. It then describes the limitations of cognitive load-based theories before reviewing state-of-the-art human mental workload measures in Ergonomics emphasising their advantages and limitations. Subsequently, it focuses on a detailed description of three self-reporting mental workload assessment techniques, these being used in the envisioned primary research study. Similarly, Cognitive Theory of Multimedia Learning and its design principles are presented in order to provide the reader with the relevant notions for the planned case study. Section 3 focuses on the design of a primary research experiment involving human learners, detailing the methodology and presenting the research hypotheses. Section 4 introduces the results of the experiment followed by a critical discussion in section 5. Section 6 concludes the paper highlighting the contribution to the body of knowledge.

2 Theoretical background

2.1 Cognitive load theory

Cognitive Load Theory (CLT) (Sweller et al., 1998) has been conceived as a form of guidance for instructional designers eager to create resources that are presented in a way that encourages the activities of the learners and optimise their performance, thus their learning (Chandler and Sweller, 1991). CLT is an approach that considers the limitations of the information processing system of the human mind (Wickens, 2008). The intuitive assumption behind this theory is that if a learner is either underloaded or overloaded, learning is likely to be adversely affected. In detail, the assumption of Cognitive Load Theory is that the capabilities of the human cognitive architecture devoted to the processing and retention of information are limited (Miller, 1956) and these limitations have a straight influence on learning. Unfortunately, the experience of mental workload is highly likely to be different on an individual basis, changing according to the learner's cognitive style, the own education and training (Paas and Van Merrienboer, 1993). As a consequence, modelling and assessing cognitive load is far from being a trivial activity. In his seminal contribution, Sweller et al. (1998) have proposed three types of cognitive load:

- intrinsic load - this is influenced by the unfamiliarity of the learners or the intrinsic complexity of the learning material under use (Ayres, 2006; Seufert et al., 2007);

- extraneous load - this is impacted by the way the instructional material is designed, organised and presented (Cierniak et al., 2009);

- germane load - this is influenced by the effort devoted for processing information, for the construction and automation of schemas in the brain of the learners (Paas and Van Merrienboer, 1993).

Intrinsic cognitive load is considered being static, extraneous load should be minimised (Mousavi et al., 1995) and germane load promoted (Debue and van de Leemput, 2014). Cognitive Load Theory, although highly relevant for instructional design and with a plethora of theoretical material that has been published in the last few decades, has a fundamental, open and challenging problem: the measurement of its three cognitive load types (De Jong, 2010; Schnotz and Kürschner, 2007; Paas et al., 2003). Unfortunately, there is little evidence that these three types are highly separable (DeLeeuw and Mayer, 2008; Sweller, 2010; Cierniak et al., 2009). Similarly, to date, there is little evidence about the ways the three different types of load can be coherently and robustly measured (Dixon, 1991; Paas et al., 2003).

According to the traditional critical rationalism proposed by Karl Popper (2014), CLT cannot be considered a scientific theory because some of its fundamental assumptions cannot be tested empirically and are thus not falsifiable (Gerjets et al., 2009). To be scientific, the measurement methods about a hypothesis must be sensitive to the different types of load. CLT must provide empirical demonstrations about the cognitive load types (its fundamental assumptions). As a consequence, the main research challenge is the development of a valid measure of cognitive load and the demonstration of the scientific value of Cognitive Load Theory and all the other theories built upon it (Goldman, 1991; Gerjets et al., 2009). CLT has mainly been developed by educational psychologists and evolved over almost three decades of research endeavour in the field of education. Despite the theoretical evolution of this theory, and the many ah-hoc, domain and context-specific applications based upon it, the practical measurement of cognitive load has not been sufficiently investigated in education. In contrast to this, the situation is different in the field of Ergonomics, where more effort has been devoted towards the development of cognitive load assessment techniques. In this discipline, cognitive load is mainly referred to as human Mental Workload (MWL), a well known psychological construct (Cain, 2007; Wickens, 2008; Young et al., 2015).

2.2 Human Mental Workload

The concept of human Mental Workload (MWL) has a long history in the fields of ergonomics and psychology, with several applications in the aviation and automotive industries. Although it has been studied for the last four decades, no clear definition of MWL has emerged that has a general validity and that is universally accepted (Cain, 2007; Longo, 2016; Rizzo et al., 2016). The main reason for assessing MWL is to measure the mental cost of performing a certain task with the goal of predicting operator and system performance (Cain, 2007). MWL is an important design criterion: at an early system design phase not only can a system or interface be optimised to take workload into consideration, but MWL can also guide designers in making appropriate structural changes (Xie and Salvendy, 2000). Modern technologies such as web applications have become increasingly complex (Longo, 2012; Longo and Dondio, 2015; Longo, 2017), with increments in the degree of MWL imposed on operators (Gwizdka, 2010; Longo, 2011). The assumption in design approaches is that as the difficulty of a task increases, perhaps due to interface complexity, MWL also increases and performance usually decreases (Cain, 2007). In turn, errors are more frequent, there are longer response times, and fewer tasks are completed per time unit. When task difficulty is negligible, systems can impose a low MWL on operators: this should be avoided as it leads to difficulties in maintaining attention and increasing reaction time (Cain, 2007). In the following sections it is shown how MWL can be measured and the formalisms to aggregate heterogeneous factors towards an overall index of mental workload. This review of current solutions is aimed at identifying both reasons why a more generally applicable measure of MWL has not yet been developed, and the key characteristics of MWL representation and assessment.

2.2.1 Measures of mental workload

The measurement of mental workload is a vast and heterogeneous topic as the related theoretical counterpart. Several assessment techniques have been proposed in the last 40 years, and researchers in applied settings have tended to prefer the use of ad hoc measures or pools of measures rather than any one measure. This tendency is reasonable, given the multi-dimensional property that characterises mental workload (Longo and Barrett, 2010; Longo, 2015; Moustafa et al., 2017). Various reviews attempted to organise the vast amount of knowledge behind MWL measures and assessment techniques (Wilson and Eggemeier, 2006; Cain, 2007; Young and Stan-

ton, 2006). In general, the measurement techniques of MWL can be classified into three broad categories:

- *self-assessment measures* including self-report measures and subjective rating scales;
- *task performance measures* which consider both primary and secondary task measures;
- *physiological measures* which are derived from the physiology of the operator.

The class of *self-report measures* is often referred to as subjective measures. This category relies on the subjective perceived experience of the interaction operator-system. Subjective measures have always appealed many workload practitioners and researchers because it is strongly believed that only the person concerned with the task can provide an accurate and precise judgement with respect to the mental workload experienced. Various dimensions and attributes of mental workload are considered in self-report measures. These include demands, performance, effort as well as individual differences such as the emotional state, attitude and motivation of the operator (Brookhuis and de Waard, 2010). The class of subjective measures include multi-dimensional approaches such as the NASA Task Load Index (Hart, 2006), the Subjective Workload Assessment Technique (Reid and Nygren, 1988), the Workload Profile (Tsang and Velazquez, 1996) as well as unidimensional approaches such as the Rating Scale Mental Effort (Zijlstra, 1993), the Subjective Workload Dominance Technique (Vidulich and Ward Frederic G., 1991) and the Bedford scale (Roscoe and Ellis, 1990). These measures and scales are mostly close-ended and, in case multidimensional, they have an aggregation strategy that combines the dimensions they are built upon to an overall index of mental workload. The class of *task performance measures* assumes that mental workload practitioners and, more generally system designers, are typically concerned with the performance of their systems and technologies. The assumption is that the mental workload of an operator, when interacting with a system, acquires importance only if it influences system performance. As a consequence, it is believed that this class of techniques is the most valuable options for designers. According to different reviews (Cain, 2007; Wilson and Eggemeier, 2006), performance measures can be classified into two sub-categories: primary task and secondary task measures. In primary-task methods the performance of the operator is monitored and analysed according to changes in primary-task demands. Examples of common measurement parameters are response and reaction time, accuracy and error rate, speed and signal detection performance, estimation time and tapping regularity. In secondary-task assessment procedures, there are two tasks involved and the performance of the secondary task may not have practical importance, but rather may serves to load or to measure the mental workload of the operator performing the primary task. The class of *physiological measures* includes bodily responses derived from the operator's physiology, and it relies on the assumption that they correlate with mental workload. They are aimed at interpreting psychological processes by analysing their effect on the state of the body, rather than measuring task performance or perceptual subjective ratings. Example includes heart rate, pupil dilation and blinking, blood pressure, brain activation signals as measured by electroencephalograms (EEG) and muscle signals as measured by electromyograms (EMG). The principal reason for adopting physiological measures is because they do not require an overt response by the operator and they can be collected continuously, within an interval of time, representing an objective way of measuring the operator state.

Subjective measures are in general easy to administer and analyse. They provide an index of overall workload and multi-dimensional measures can determine the source of mental workload. However, the main drawback is that they can only be administered post-task, thus influencing the reliability for long tasks. In addition, meta-cognitive limitations can diminish the accuracy of reporting and it is difficult to perform comparisons among raters on an absolute scale. However, they appear to be the most appropriate types of measurement for assessing mental workload because they have demonstrated high levels of sensitivity and diagnosticity (Rubio et al., 2004). *Task performance measures* can be primary or secondary. Primary-task measures represent a direct index of performance and they are accurate in measuring long periods of mental workload. They are capable of discriminating individual differences in resource competition. However, the main limitation is that they cannot distinguish performance of multiple tasks that are executed simultaneously by an operator. If taken in isolation, they do not represent reliable measures, though if used in conjunction with other measures, such as subjective ratings, they can be useful. Secondary task measures have the capacity of discriminating between tasks when no differences are detected in primary performance. They are useful for quantifying the individual's spare attentional capacity as well as short periods of workload. However, they are only sensitive to large changes in mental workload and they might be highly intrusive, influencing the behaviours of users while interacting with the primary task. *Physiological measures* are extremely good at

monitoring data on a continuous interval, thus having high measurement sensitivity. They do not interfere with the performance on the primary task. However, the main drawback is that they can be easily confounded by external interference. Moreover, they require equipment and tools that are often physically obtrusive and the analysis of data is complex, requiring well trained experts. In the experimental study carried out in this research, subjective mental workload measures have been adopted because they are easy to be administered in a typical third-level classroom. Primary and secondary task measures would have been intrusive and would have influenced the natural behaviour of learners in the classroom. Physiological measures would have been physically obtrusive, requiring expensive equipment to be attached to the body of each learner. The next sections describe the three MWL assessment techniques adopted in the current study, describing their formalism to produce a quantifiable score of mental workload.

2.3 Subjective workload techniques

The NASA Task Load Index (NASATLX) instrument (Hart, 2006) belongs to the category of self-assessment measures. It has been validated in the aviation industry and in other contexts within Ergonomics (Hart, 2006; Rubio et al., 2004) with several applications in many socio-technical domains. It is a combination of six factors believed to influence MWL. Each factors is quantified with a subjective judgement coupled with a weight computed via a paired comparison procedure. Subjects are required to decide, for each possible pair (binomial coefficient, $\binom{6}{2} = 15$) of the 6 factors, *'which of the two contributed the most to mental workload during the task'*, such as 'Mental or Temporal Demand?', and so forth. The weights w are the number of times each dimension was selected. In this case, the range is from 0 (not relevant) to 5 (more important than any other attribute). The final MWL score is computed as a weighed average, considering the subjective rating of each attribute d_i and the correspondent weights w_i (averaged here, and scaled in $[1..100] \in \Re$ for comparison purposes - equation 1). For the NASA-TLX questionnaire we refer the reader to (Longo, 2017).

$$NASATLX : [0..100] \in \Re \quad = \quad \left(\sum_{i=1}^{6} d_i \times w_i \right) \frac{1}{15} \quad (1)$$

The Workload Profile (WP) assessment procedure (Tsang and Velazquez, 1996) is built upon the Multiple Resource Theory proposed in Wickens (2008). In this theory, individuals are seen as having different capacities or 'resources' related to:

- *stage of information processing* – perceptual/central processing and response selection/execution;
- *code of information processing* – spatial/verbal;
- *input* – visual and auditory processing;
- *output* – manual and speech output.

Each dimension is quantified through subjective rates and subjects, after task completion, are required to rate the proportion of attentional resources used for performing it with a value in the range $0..1 \in \Re$. A rating of 0 means that the task placed no demand while 1 indicates that it required maximum attention. The aggregation strategy is a simple sum of the 8 rates d (averaged here, and scaled in $[1..100] \in \Re$ for comparison purposes - equation 2). For details about the questionnaire associated to it we refer the reader to (Longo, 2017).

$$WP : [0..100] \in \Re \qquad WP = \frac{1}{8} \sum_{i=1}^{8} d_i \times 100 \quad (2)$$

The Rating Scale Mental Effort (RSME) is a unidimensional procedure that considers the exerted subject's effort, and subjective ratings are indicated across a continuous line, within the interval 0 to 150 with ticks each 10 units (scale 3). Labels such as 'absolutely no effort', 'considerable effort' and 'extreme effort' are used along the line. The final mental workload of a subject is related to the exerted effort indicated on the line by the subject, from the origin of the scale (zero). Although the procedure is relatively simple and quick, it has showed a good degree of sensitivity. However, on the other hand, it has demonstrated to have a poor diagnostic capacity (Zijlstra, 1993). For details about the scale, its history, and development, we refer the reader to Zijlstra (1993).

$$RSME : [0..150] \in \Re \quad (3)$$

2.4 Cognitive Theory of Multimedia Learning

Another popular cognitivist theory of learning is the Cognitive Theory of Multimedia Learning (CTML), proposed by Prof. Mayer (2002, 2017). This theory is strictly supported by other learning theories, including Sweller's theory of Cognitive Load. CTML is based upon three assumptions:

- dual-channel - two separate channels exist for processing information in the human brain, namely the auditory and the visual channel; this assumption has been inspired by the dual-coding approach of Paivio (1990);

- limited processing capacity - each channel has a finite, limited capacity; this is in line with the assumption of Cognitive Load Theory (Sweller et al., 1998) and aligned to Baddeley's models of working memory (Baddeley and Hitch, 1974);
- active processing - learning is an active process that includes the selection, the filtering, the organisation of information and the integration of this to prior knowledge

Humans can process a finite amount of information in each channel at a time. In details, according to the Cognitive Theory of Multimedia Learning, the human brain does not interpret multimedia instructions made by words, auditory information and pictures in a mutually exclusive way. On the contrary, all these forms of information are firstly selected and then organised dynamically to produce mental logical representations (schemas). These are particular cognitive constructs able to organise information for storage in long-term memory. In details, schemas are capable of organising simpler elements in a way these can subsequently act as elements in higher-order schemas. Learning coincides with the development of complex schema and the transferring of those procedures that are learned from controlled processing to automated processing. This shift frees working memory that can be used for other cognitive processes. Mayer (2005) suggested five ways of representing words and pictures while information is processed in memory. These are particular stages of processing information. The first is the words and pictures in the multimedia presentation layer. The second form includes the acoustic (sounds) and iconic representation (images) in sensory memory. The third form concerns the sounds and images within working memory. The fourth form coincides with the verbal and pictorial models, always within working memory. The fifth form relates prior knowledge, or schemas, stored in long-term memory. In relation to instructional design, Mayer proposed a set of principles for creating instructions aligned to the above limitations of the brain and the dual-channel paradigm of learning. Readers can obtain more information on the principles in Mayer (2009). Generally speaking, these design principles suggest to provide learners with coherent instructional material in the form of verbal and pictorial information. Coherent information is aimed at guiding the learners to select the relevant words and pictures therefore reducing the cognitive load in each elicited channel. CTML is strictly connected to the Cognitive Load Theory because its twelve principles can be grouped according to the three types of loads - reducing extraneous load: coherence, signaling, redundancy, spatial contiguity, temporal contiguity; managing intrinsic load - segmenting, pretraining, modality fostering; germane load - multimedia, personalisation, voice, image. These principles have emerged from more than 100 studies conducted in the field (Mayer, 2009). In addition to these, advanced principles have been proposed by Mayer in a number of papers, and recently updated (Mayer, 2017). This demonstrates how his theory is a dynamic one, suggesting how the principles should not be taken rigidly, but rather as a starting point for discussion and experimentation. Cognitive Theory of Multimedia Learning has been described for providing the readers with those key elements necessary for the comprehension of the primary research experiment presented in the next section.

3 Design, methodology, hypotheses

A primary research experiment has been designed to investigate the reliability and the validity of the three aforementioned subjective mental workload assessment techniques (NASA, WP, RSME) as well as their sensitivity to discriminate different design conditions. An experiment has been conducted in the School of Computing at the Dublin Institute of Technology, Ireland, in the context of an MSc module: 'Research design and proposal writing'. This module is usually taught both to full-time and part-time students. The main difference between full-timers and part-timers is the way classes are planned for them. Full-timers attend 12 classes within an academic semester, of 2 hours each, on a day of the week. Part-timers attend 4 classes of 6 hours, within an academic semester. Each class is scheduled on a Saturday and are usually separated by a period of 3 to 4 weeks of inactivity. Full-timers have usually no break during their classes, while part-timers, given the long day in class, have two to three breaks (coffees and lunch). In this primary research, the part-time cohort has been chosen, and only the first class (out of four) has been selected. Four topics were presented to part-timers during the first class (Saturday): 'Science', 'The Scientific Method' 'Planning Research' and 'Literature Review'. The subsequents classes were focused on more practical and collaborative activities where students had to put in practice theoretical notions. The rationale behind the selection of the part-time cohort and the first class are various. The first reason is due to the nature of the taught subject: theoretical at the beginning of the semester and more practical towards its end. This would have allowed the delivery of the four topics, during the first class, in a controlled one-way style, from the lecturer to the stu-

Table 1: Comparison of design conditions according to the principles of Cognitive Theory of Multimedia Learning

Principle	CLT load type	Design condition (A)	Design condition (B)
coherence	extraneous	any extraneous material was kept to minimum.	
signaling	extraneous	cues, in the form of relevant keywords, with a larger font size	cues (relevant keywords), popped-in in the video to emphasise the organisation of essential material.
redundancy	extraneous	graphical aids and use of narratives	most of text was removed, offloading one channel (eyes); graphical aids and the use of narratives.
spatial contiguity	extraneous	corresponding words and pictures were placed beside each other and not in different slides or screens.	
temporal contiguity	extraneous	corresponding words and pictures were presented at the same time	corresponding words (verbally transmitted) and pictures were presented at the same time.
segmenting	intrinsic	the instructional material was presented in a single unit	the instructional material is presented in segments, separated by video transitions.
pre-training	intrinsic	no pre-training was offered to students.	
modality	intrinsic	printed text is kept in the slides and verbally explained	printed text is removed, offloading one channel (eyes) and verbally explained (ears.)
multimedia	germane	words and pictures.	
personalisation	germane	words are presented using a conversational style and not a formal style	
voice	germane	the words are spoken by the lecturer and not by an artificial machine voice.	
image	germane	no video was used, thus no speaker's image was available	the lecturer's image was most of the time kept in the video, sometimes using the full space available or using half-space, with the second half used for important pieces of text or pictures. Other times, the image was removed and important sentences were textually presented in the full screen.

dents. In other words, this would have facilitated the application of the three subjective mental workload assessment techniques - the NASA-TLX, the Workload Profile and the Rating Scale Mental Effort - at the end of the delivery of each topic, without interruptions. The second reason lies in the ease of manipulation of this traditional one-way delivery method without altering the content of each topic. In fact, by keeping the content constant, a number of delivery methods could have been employed, including for instance, a verbal presentation of the content backed up with a set of slides projected on a white board; a verbal presentation of the content with relevant keywords written on a black-board; a verbal presentation of the content supported by diagrams; a multimedia presentation making use of pictorial and acoustic material and many others. The third reason refers to the state of mind of each individual learner during the long class. In fact, students were expected to loose interest during the day, with a constant reduction of their engagement and the effort exerted towards learning. All these factors along with other individual characteristics of each learner were expected to increase the overall cognitive load towards the upper limit, due to fatigue, or to decrease it towards the lower limit, due to boredom. For experimental purposes, and taking into account the above rationale, two design conditions were eventually formed. These conditions were built according to the design principles of the Cognitive Theory of Multimedia Learning (CTML) - as described in section 2.4. In detail, the differences between the two design conditions are described in table 1, grouped by the underpinning principles of the CTML. Figure 2 summarises the full research design.

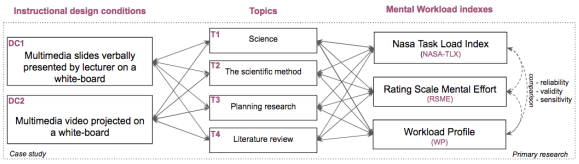

Figure 2: Layout of the design of the experiment

Research hypotheses Informally, the research hypotheses are that the NASA Task Load Index, the Workload Profile and the Rating scale mental effort are reliable and valid measures of mental workload when applied in an educational context. If this will be

Table 2: Criteria for the evaluation of different mental workload assessment techniques, their definition, associated statistical tests and the expectations for this primary research

Criteria	Definition	Associated test	expectation
Reliability	the consistency or stability of a MWL measurement	Cronbach's Alpha	high
Validity (face)	the extent to which a MWL measurement is subjectively viewed as covering the construct of MWL itself	Pearson/ Spearman correlation	positive & moderate
Validity (convergent)	the degree to which two measures of MWL, expected to be theoretically related, are in fact related	Pearson/ Spearman correlation	positive & moderate
Sensitivity	the extent to which a MWL measurement is able to detect changes in instructional design condition	ANOVA + T-test/ Wilcoxon test	moderate

the case, then the extent to which these instruments can discriminate the two design conditions will be investigated by computing a measure of their sensitivity. Table 2 formally presents the hypotheses, listing the criteria for evaluating the selected mental workload measures, their definition, the associated statistical test and the expected outcome. Note that both forms of validity are expected to be moderate. A high degree of face validity would imply that participants could subjectively and precisely assess the construct of mental workload as good as the selected mental workload measures. Therefore these measures would not have reason to exist as participants coul precisely assess mental workload autonomously. Similarly, a high degree of convergent validity would imply that two different measures assess the construct of mental workload exactly in the same way, but given the known difficulties in measuring this construct, the chances that this occurs are low. As a consequence a positive moderate correlation is expected for both the forms of validity, underlying a reasonable relationship of the selected mental workload measures.

3.1 Participants and procedure

Two different groups of part-time students participated in the experimental study and attended the long-day of the MSc module 'Research design and proposal writing' in two different semesters. Both the groups attended the four topics listed in figure 2 in the same order (T1-T4). The first group received the first design condition (DC1) while the second group received the second design condition (DC2). At the end of each topic, students were asked to fill questionnaires in, aimed at quantifying the mental workload experienced during the class. The NASA-TLX and the WP are multi-dimensional and thus require participants to answer a number of questions. To facilitate the completion of each questionnaire and not to overload students with many questions, two sub-groups were formed, one receiving the NASA-TLX and one the WP. Eventually, both the groups received the RSME questionnaire. The rationale was that, being RSME uni-dimensional, adding one further question to the previous questionnaires was deemed reasonable. In summary, the two subgroups are:

- sub-group A: the NASA-TLX + the RSME
- sub-group B: the WP + the RSME

Students were instructed about the study and were required to sign a consent form. This documentation was approved by the ethics committee of the Dublin Institute of Technology. Students had the right to withdrawn at any time during the experiment and collection of data. The formation of the two subgroups was random for each topic, therefore students could receive any questionnaire at any given time. Table 3 summarises the groups and sub-groups formed, aggregated by topic and the design condition received. It also lists the number of students who participated, and the length of each topic. Note that some of the student who took part in the experimental study did not fully complete the administered questionnaires, therefore associated data was discarded. Additionally, due to the fact that each class was rather long (7 hours), some student left the classroom at some stage. As a consequence, the number of people who attended a topic within the day was not the same across topics.

4 Results

Table 4 presents the descriptive statistics of each subgroup introduced in table 3. In details, it shows the average (avg), the standard deviation (std), the median (med) and the Shapiro-Wilk test (W) of normality of the distributions, along its p-value (p-val), of the mental workload scores obtained across the different topics and the mental workload assessment techniques (NASA, WP, RSME), grouped by design condition

Table 3: Descriptions of topics, design condition, groups, workload instruments received and number of students per group

Topic	Condition (group)	sub-groups (# of students)		Length (mins)
		A (RSME+NASA)	B (RSME+WP)	
T1 - Science	DC1	8	11	62.00
T2 - The scientific method	DC1	10	13	46.00
T3 - Planning research	DC1	11	9	54.00
T4 - Literature Review	DC1	11	9	41.00
T1 - Science	DC2	13	13	17.24
T2 - The scientific method	DC2	12	12	27.50
T3 - Planning research	DC2	11	11	10.34
T4 - Literature Review	DC2	13	11	18.14

Table 4: Descriptions of topics, design condition received, mental workload questionnaires administered and descriptive statistics for each subgroup (average, standard deviation, median, Shapiro-Wilk test (W) of normality with p-value and 95% confidence level)

Topic	Condition	Mental Workload assessment technique											
		NASA				WP				RSME			
		avg	std	med	W/p-val	avg	std	med	W/p-val	avg	std	med	W/p-val
T1	DC1	45.0	09.0	45.0	0.95 / 0.69	55.9	20.6	53.1	0.94 / 0.54	45.6	23.9	40.0	0.88 / 0.03
T2	DC1	54.3	11.6	54.0	0.94 / 0.54	51.0	16.5	51.8	0.95 / 0.54	59.7	25.9	60.0	0.96 / 0.57
T3	DC1	50.2	12.8	53.6	0.90 / 0.25	50.2	15.9	53.1	0.91 / 0.29	54.9	20.8	51.5	0.90 / 0.04
T4	DC1	46.0	13.6	49.6	0.96 / 0.78	52.1	5.60	53.7	0.91 / 0.30	56.7	21.2	55.0	0.95 / 0.30
T1	DC2	40.8	17.1	37.3	0.89 / 0.11	42.4	14.9	38.7	0.94 / 0.44	43.6	19 .0	40.0	0.90 / 0.01
T2	DC2	49.4	10.4	48.1	0.98 / 0.99	55.0	09.5	54.0	0.93 / 0.35	61.4	19.0	62.5	0.93 / 0.09
T3	DC2	47.3	13.0	50.0	0.97 / 0.90	43.5	13.7	43.1	0.98 / 0.94	47.9	18.3	47.5	0.93 / 0.14
T4	DC2	52.2	16.4	48.3	0.96 / 0.74	45.5	19.2	44.3	0.90 / 0.17	59.0	19.0	52.5	0.91 / 0.04

(DC1, DC2) and topic (T1-T4). As it is possible to assess from table 4, most of the p-values (p-val) or the Shapiro-Wilk test (W) are greater than the chosen alpha level ($\alpha = 0.05$), thus for most of the subgroups, the null hypothesis that the data came from a normally distributed population cannot be rejected (is accepted). As a consequence, most of the MWL scores across the topics follow a normal distribution.

Reliability To assess the reliability of the selected mental workload instruments, Cronbach's Alpha has been employed. It measures the internal consistency of the items of a multi-dimensional instrument, that means, how closely related these items are as a group. For this reason, the Rating Scale Mental Effort is not subject to reliability analysis as it is uni-dimensinal. Table 5 shows the Cronbach's Alpha coefficients of the other two selected multidimensional mental workload assessment instruments, namely the NASA-TLX and the Workload Profile, obtained by considering all the answers of students across all the topics and design conditions. In most sciences, a reliability coefficient of .70 or higher is considered acceptable to infer that a scale is a consistent measure of a construct. Therefore, both the Nasa Task Load Index and the Workload Profile can be considered reliable measures of mental workload, as assessed with the data collected in this primary research. To confirm the obtained high reliability, Cronbach's Alpha has been computed also for each topic and design condition. Table 6 demonstrates how the reliability scores are mostly above 0.7 across the topics and design conditions. Therefore there is a strong evidence suggesting how the NASA-TLX and Workload Profile might be reliably applied in educational contexts.

Table 5: Reliability of the multidimensional mental workload scales with sample size, related number of items in the scales and associated Cronbach's Alpha

Instrument	Sample size	# of items	Cronbach's α
NASA	89	6	0.75
WP	89	8	0.87

Table 6: Reliability of the multidimensional mental workload scales, namely the Nasa Task Load Index and the Workload Profile, grouped by topic

Topic	Condition	Mental Workload Instruments			
		NASA-TLX		WP	
	(group)	Size	C's α	Size	C's α
T1	DC1	8	0.72	11	0.94
T2	DC1	10	0.68	13	0.89
T3	DC1	11	0.59	9	0.93
T4	DC1	11	0.86	9	0.23
T1	DC2	13	0.85	13	0.82
T2	DC2	12	0.45	12	0.6
T3	DC2	11	0.76	11	0.83
T4	DC2	13	0.81	11	0.92

Validity To assess the validity of the three selected MWL assessment instruments, two sub-forms of validity were selected, namely face and convergent validity. The former measures the extent to which a MWL measurement is subjectively viewed as covering the construct of MWL itself while the latter measures the degree to which two measures of MWL, expected to be theoretically related, are in fact related. To assess face validity, a question of overall MWL has been asked to students after the completion of each topic (figure 3) and before the completion of the MWL questionnaires (NASA/WP). Answers to this new question have been correlated to the MWL scores of the other MWL techniques (NASA/WP/RSME). To assess convergent validity, the MWL scores produced by the multidimensional NASA-TLX and the WP instruments have been correlated against the MWL scores produced by the unidimensional RSME instrument. Note that this was possible because a participant filled in the questionnaire associated to the NASA-TLX or WP and the RSME. Correlation between the NASA-TLX and WP cannot be computed because no participant received both the questionnaires associated to these instruments at the same time. Both the Pearson and the Spearman's Rank correlation coefficients have been employed for computing validity. Tables 7, 8 respectively shows the correlations for face validity and convergent validity.

Table 8: Convergent validity of the mental workload assessment instruments, sample size, Pearson and Spearman correlation coefficients

Instrument	size	Pearson r	Spearman ρ
NASA-TLX vs RSME	89	0.45	0.43
WP vs RSME	89	0.40	0.48

Figure 3: Question for face validity detection

How much mental workload the teaching session imposed on you?

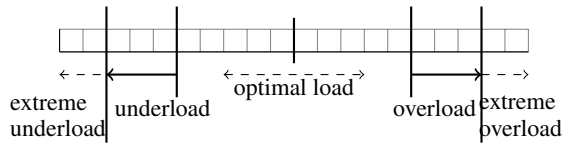

Table 7: Face validity of the mental workload assessment instruments, sample size, Pearson and Spearman correlation coefficients

Instrument	Sample size	Pearson r	Spearman ρ
NASA	89	0.57	0.61
WP	89	0.51	0.56
RSME	178	0.44	0.42

Sensitivity The sensitivity of the selected MWL instruments has been calculated performing and analysis of the variance of their MWL fscores. A formal comparison has been carried out to check whether the distributions of the MWL scores for each topic are statistically significant different across the two design conditions (table 9). Independent two-sample T-Tests (t) have been adopted in most of the cases, when the two underlying distributions are normal, while the Wilcoxon signed-rank test (V) when distributions are not normal. In table 9, all the p-values associated to the T-tests are greater than 0.05, therefore it is possible to conclude that the means of the two groups under comparison are significantly similar. Similarly, since the p-values associated to the V-tests are greater than 0.05, it is possible to conclude that the means have remained essentially unchanged. These findings confirm that there is no difference between the first design condition and the second design condition across the four topics in terms of mental workload variation.

5 Discussion

Two multidimensional and a unidimensional subjective mental workload (MWL) assessment technique, borrowed from the discipline of Ergonomics, have been employed in a novel primary research experiment within Education. The former are the Nasa Task Load Index (Hart, 2006) and the Workload Profile (Tsang and Velazquez, 1996) while the latter is the Rating Scale Mental Effort (Zijlstra, 1993). These instruments have been applied in a typical third-level classroom in the context of a module taught in the School of Computing, to part-time master students, at

Table 9: Comparison of distributions of the workload scores using t-test and Wilcoxon-test at 95% confidence level

Topic	NASA				WP				RSME			
	t	p-val	V	p-val	*t*	p-val	*V*	p-val	*t*	p-val	*V*	p-val
T1	0.64	0.53	69	0.22	1.86	0.08	104	0.06	0.32	0.75	252.5	0.9
T2	1.04	0.31	74	0.36	-0.74	0.47	61.5	0.37	-0.26	0.8	271	0.91
T3	0.52	0.61	69.5	0.55	1	0.33	64	0.29	1.15	0.26	263	0.27
T4	-0.99	0.33	55.5	0.35	0.99	0.34	66.5	0.2	-0.38	0.71	223	0.69

the Dublin Institute of Technology. The experiment involved the quantification of the experienced mental workload of two groups of part-time students who were exposed to two different design conditions of the same topics. The former condition included the delivery of four topics by employing a traditional lecturer-students delivery of instructional material employing slides projected to a white-board built with text, pictures and diagrams. The latter condition included the delivery of the same four topics through multimedia video presentations built by following a set of principles of the Cognitive Theory of Multimedia Learning Mayer (2009). An analysis of the reliability of the two multidimensional MWL assessment techniques has been conducted through a measure of their internal consistency. In details, Cronbach's Alpha has been employed to assess the relation of the items associated to each technique. An obtained alpha value of 0.75 for the NASA task Load Index suggested that all its items share high covariance and probably measure the underlying construct (mental workload). The situation is similar for the Workload Profile with an even higher alpha of 0.87. Although the standards for what can be considered a 'good' alpha coefficient are entirely arbitrary and depend on the theoretical knowledge of the scales in question, results are in line with what literature recommends: a minimum coefficient between 0.65 and 0.8 is required for reliability. Having reliable multidimensional measures of mental workload, an analysis of their validity has been subsequently performed. In detail, two forms of validity were assessed: face and convergent validity. The former validity indicates the extent to which the three employed MWL measures - the Nasa Task Load Index, the Workload Profile and the Rating Scale Mental Effort - are subjectively viewed as covering the construct of MWL itself by subjects. The latter validity indicates the degree to which the two multidimensional measures of MWL are theoretically related to the unidimensional measure. The obtained Pearson and Spearman coefficients suggest how the three MWL measures are moderately correlated to the indication of overall MWL self-reported by subjects, thus demonstrating moderate face validity. Similarly, correlation coefficients show the moderate relationship that exist between the two multidimensional MWL measures and the unidimensional MWL measure, thus demonstrating moderate convergent validity. Eventually, with the expected moderate validity, the sensitivity of the three measures of MWL was subsequently computed. Sensitivity referred to the extent to which a MWL measure was able to detect changes in instructional design conditions. In detail, sensitivity was assessed through an analysis of the variance of the MWL scores associated to the four topics across the two design conditions with a formal comparison of their distributions using the T-test or the Wilcoxon test. Evidence strongly suggests how the two design conditions imposed on average similar mental workload to students as computed by the three MWL assessment techniques. Eventually, given the strong reliability and moderate validity achieved by these techniques, a reasonable conclusion is that the design principles from the Cognitive Theory of Multimedia Learning (CTML)- used to design the second condition - were, in this primary research, as not as effective as expected. Future work might include the application of more advanced principles of CTML (Mayer, 2005) to develop additional design conditions. This might include the application of the navigation principle by which humans learn better in environments where appropriate navigational aids are provided or the collaborative principle by which people learn better when involved in collaborative activities.

6 Conclusions

This study attempted to investigate the impact of three mental workload (MWL) assessment techniques, namely the NASA Task Load Index, the Workload Profile and the Rating Scale Mental Effort, for the evaluation of different instructional design conditions. A primary research study has been performed in a typical third-level classroom and a case study involved the consideration of two design conditions. The former condition included the delivery of four topics by employing a traditional lecturer-students delivery of instructional material employing textual and pictorial slides projected

to a white-board, including diagrams. The latter condition included the delivery of the same content through multimedia videos built by employing a set of principles from Cognitive Theory of Multimedia Learning (Mayer, 2009). Evidence strongly suggests how the three MWL measures are reliable when applied to a typical third-level classroom. Results demonstrated their moderate validity, in line with the validity achieved in other experiments within Ergonomics. On the contrary, their sensitivity was very low in discriminating the two design conditions. However, given the high reliability and modest validity of the three MWL measures, the achieved sensitivity might reasonably underlines the minimal impact of the principles of Cognitive Theory of Multimedia Learning for developing the second design condition and alter the experienced mental workload by learners. The contributions of this research are to offer a new perspective on the application of mental workload measures within the field of Education, and a richer approach to support instructional design. Additionally, contrarily to the lack of falsifiability of Cognitive Load Theory and its load types, as emerged in the literature, this study conforms to the Popperian's view of science, this being replicable and falsifiable. Every single test of existing methods of mental workload assessment in Education is aimed at increasing our understanding and the ways this construct can be applied for instructional design.

REFERENCES

Ayres, P. (2006). Using subjective measures to detect variations of intrinsic cognitive load within problems. *Learning and Instruction*, 16(5):389–400.

Baddeley, A. and Hitch, G. (1974). *Working memory*, volume 8, pages 47–90. Academic Press.

Brookhuis, K. A. and de Waard, D. (2010). Monitoring drivers' mental workload in driving simulators using physiological measures. *Accident Analysis & Prevention*, 42(3):898–903.

Cain, B. (2007). A review of the mental workload literature. Technical report, Defence Research and Development Canada Toronto.

Chandler, P. and Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction*, 8(4):293–332.

Cierniak, G., Scheiter, K., and Gerjets, P. (2009). Explaining the split-attention effect: Is the reduction of extraneous cognitive load accompanied by an increase in germane cognitive load? *Computers in Human Behavior*, 25(2):315–324.

De Jong, T. (2010). Cognitive load theory, educational research, and instructional design: some food for thought. *Instructional science*, 38(2):105–134.

Debue, N. and van de Leemput, C. (2014). What does germane load mean? an empirical contribution to the cognitive load theory. *Frontiers in Psychology*, 5:1099.

DeLeeuw, K. E. and Mayer, R. E. (2008). A comparison of three measures of cognitive load: Evidence for separable measures of intrinsic, extraneous, and germane load. *Journal of Educational Psychology*, 100(1):223.

Dixon, P. (1991). From research to theory to practice: Commentary on chandler and sweller. *Cognition and Instruction*, 8(4):343–350.

Gerjets, P., Scheiter, K., and Cierniak, G. (2009). The scientific value of cognitive load theory: A research agenda based on the structuralist view of theories. *Educational Psychology Review*, 21(1):43–54.

Goldman, S. R. (1991). On the derivation of instructional applications from cognitive theories: Commentary on chandler and sweller. *Cognition and Instruction*, 8(4):333–342.

Gwizdka, J. (2010). Distribution of cognitive load in web search. *Journal of the american society & information science & technology*, 61(11):2167–2187.

Hart, S. G. (2006). Nasa-task load index (nasa-tlx); 20 years later. In *Human Factors and Ergonomics Society Annual Meeting*, volume 50, pages 904–908, San Francisco, California, USA. Sage Journals.

Longo, L. (2011). Human-computer interaction and human mental workload: Assessing cognitive engagement in the world wide web. In *INTERACT (4)*, pages 402–405.

Longo, L. (2012). Formalising human mental workload as non-monotonic concept for adaptive and personalised web-design. In *User Modeling, Adaptation, and Personalization*, volume 7379, pages 369–373. Springer.

Longo, L. (2015). Designing medical interactive systems via assessment of human mental workload. In *Int. Symposium on Computer-Based Medical Systems*, pages 364–365.

Longo, L. (2016). Mental workload in medicine: Foundations, applications, open problems, challenges and future perspectives. In *2016 IEEE 29th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 106–111.

Longo, L. (2017). Subjective usability, mental workload assessments and their impact on objective human performance. In *IFIP Conference on Human-Computer Interaction*, pages 202–223. Springer.

Longo, L. and Barrett, S. (2010). A computational analysis of cognitive effort. In *Intelligent Information and Database Systems, Second International Conference, ACIIDS, Hue City, Vietnam*, volume LNCS 5991 of *Lecture Notes in Computer Science*, pages 65–74. Springer.

Longo, L. and Dondio, P. (2015). On the relationship between perception of usability and subjective mental workload of web interfaces. In *IEEE/WIC/ACM Int. Conf. on Web Intelligence and Intelligent Agent Technology, WI-IAT 2015, Singapore, December 6-9, Volume I*, pages 345–352.

Mayer, R. (2017). Using multimedia for e-learning. *Journal of Computer Assisted Learning*, 33(5):403–423. JCAL-16-266.R1.

Mayer, R. E. (2002). Multimedia learning. *Psychology of Learning and Motivation*, 41:85–139.

Mayer, R. E. (2005). *The Cambridge handbook of multimedia learning*. Cambridge university press.

Mayer, R. E. (2009). *Multimedia learning*. Cambridge University Press.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63(2):81–97.

Mousavi, S., Low, R., and Sweller, J. (1995). Reducing cognitive load by mixing auditory and visual presentation modes. *Journal of Educational Psychology*, 87(2):319–334.

Moustafa, K., Luz, S., and Longo, L. (2017). Assessment of mental workload: a comparison of machine learning methods and subjective assessment techniques. In *Int. Symposium on Human Mental Workload: Models and Applications*, pages 30–50.

Paas, F., Tuovinen, J. E., Tabbers, H., and Van Gerven, P. W. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational psychologist*, 38(1):63–71.

Paas, F. and Van Merrienboer, J. J. G. (1993). The efficiency of instructional conditions: An approach to combine mental effort and performance measures. *Human Factors: the Journal of the Human Factors and Ergonomics Society*, 35(4):737–743.

Paivio, A. (1990). *Mental Representations: A Dual Coding Approach*. Oxford Psychology Series. Oxford University Press.

Popper, K. (2014). *Conjectures and refutations: The growth of scientific knowledge*. routledge.

Reid, G. B. and Nygren, T. E. (1988). The subjective workload assessment technique: A scaling procedure for measuring mental workload. In Hancock, P. A. and Meshkati, N., editors, *Human Mental Workload*, volume 52 of *Advances in Psychology*, chapter 8, pages 185–218. North-Holland.

Rizzo, L., Dondio, P., Delany, S. J., and Longo, L. (2016). *Modeling Mental Workload Via Rule-Based Expert System: A Comparison with NASA-TLX and Workload Profile*, pages 215–229. Springer International Publishing, Cham.

Roscoe, A. H. and Ellis, G. A. (1990). A subjective rating scale for assessing pilot workload in flight: A decade of practical use. Technical report TR 90019, Royal Aerospace Establishment.

Rubio, S., Diaz, E., Martin, J., and Puente, J. M. (2004). Evaluation of subjective mental workload: A comparison of swat, nasa-tlx, and workload profile methods. *Applied Psychology*, 53(1):61–86.

Schnotz, W. and Kürschner, C. (2007). A reconsideration of cognitive load theory. *Educational Psychology Review*, 19(4):469–508.

Seufert, T., Jänen, I., and Brünken, R. (2007). The impact of intrinsic cognitive load on the effectiveness of graphical help for coherence formation. *Computers in Human Behavior*, 23(3):1055–1071.

Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational psychology review*, 22(2):123–138.

Sweller, J., Van Merrienboer, J., and Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, 10(3):251–296.

Tsang, P. S. and Velazquez, V. L. (1996). Diagnosticity and multidimensional subjective workload ratings. *Ergonomics*, 39(3):358–381.

Vidulich, M. A. and Ward Frederic G., S. J. (1991). Using the subjective workload dominance (sword) technique for projective workload assessment. *Human Factors Society*, 33(6):677–691.

Wickens, C. D. (2008). Multiple resources and mental workload. *Human Factors*, 50(2):449–454.

Wilson, G. F. and Eggemeier, T. F. (2006). Mental workload measurement. In Karwowski, W., editor, *Int. Encyclopedia of Ergonomics and Human Factors (2nd ed.)*, volume 1, chapter 167. Taylor and Francis.

Xie, B. and Salvendy, G. (2000). Review and reappraisal of modelling and predicting mental workload in single and multi-task environments. *Work and Stress*, 14(1):74–99.

Young, M. S., Brookhuis, K. A., Wickens, C. D., and Hancock, P. A. (2015). State of science: mental workload in ergonomics. *Ergonomics*, 58(1):1–17.

Young, M. S. and Stanton, N. A. (2006). Mental workload: theory, measurement, and application. In Karwowski, W., editor, *Encyclopedia of ergonomics and human factors*, volume 1, pages 818–821. Taylor & Francis, 2nd edition.

Zijlstra, F. R. H. (1993). *Efficiency in work behaviour*. Doctoral thesis, Delft University, The Netherlands.