Technological University Dublin
ARROW@TU Dublin

Conference papers

School of Computer Sciences

2019

An Evaluation of the Reliability, Validity and Sensitivity of Three Human Mental Workload Measures Under Different Instructional Conditions in Third-Level Education

Luca Longo

Giuliano Orru

Follow this and additional works at: https://arrow.tudublin.ie/scschcomcon

Part of the Computer Sciences Commons, Educational Technology Commons, and the Other Education Commons

This Conference Paper is brought to you for free and open access by the School of Computer Sciences at ARROW@TU Dublin. It has been accepted for inclusion in Conference papers by an authorized administrator of ARROW@TU Dublin. For more information, please contact arrow.admin@tudublin.ie, aisling.coyne@tudublin.ie, gerard.connolly@tudublin.ie.

This work is licensed under a Creative Commons Attribution-Noncommercial-Share Alike 4.0 License

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/333890330

An Evaluation of the Reliability, Validity and Sensitivity of Three Human Mental Workload Measures Under Different Instructional Conditions in Third-Level Education

Conference Paper · June 2019

DOI: 10.1007/978-3-030-21151-6_19

CITATIONS
11

READS
288

2 authors:

Luca Longo
Technological University Dublin - City Campus
90 PUBLICATIONS **1,082** CITATIONS
SEE PROFILE

Giuliano Orru
Technological University Dublin - City Campus
6 PUBLICATIONS **78** CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Project Community of inquiry and Cognitive load Theory View project

All content following this page was uploaded by Luca Longo on 23 July 2019.

The user has requested enhancement of the downloaded file.

An evaluation of the reliability, validity and sensitivity of three human mental workload measures under different instructional conditions in third-level education

Luca Longo[1], 2, Giuliano Orru[1]

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

Abstract. Although Cognitive Load Theory (CLT) has been researched for many years, it has been criticised for its theoretical clarity and its methodological approach. A crucial issue is the measurement of three types of cognitive load conceived in the theory, and the assessment of overall human cognitive load during learning tasks. This research study is motivated by these issues and it aims to investigate the reliability, validity and sensitivity of three existing self-reporting mental workload instruments, mainly used in Ergonomics, when applied to Education and in particular to the field of Teaching and Learning. A primary research study has been designed and performed in a typical third-level classroom in Computer Science, and the self-reporting mental workload instruments employed are the NASA Task Load Index, the Workload Profile and the Rating Scale Mental Effort. Three instructional design conditions have been designed and employed for the above purposes. The first design condition followed the traditional explicit instruction paradigm whereby a lecturer delivers instructional material mainly using a one-way approach with almost no interactions with students. The second design condition was inspired by the Cognitive Theory of Multimedia Learning whereby the same content, delivered under the first condition, was converted in a multimedia video by following a set of its design principles. The third design condition was an extension of the second condition whereby an inquiry activity was executed after the delivery of the second condition. The empirical evidence gathered in this study suggests that the three selected mental workload measures are highly reliable. Their moderate face validity is in line with the results obtained so far within Ergonomics emphasising and confirming the difficulty in creating optimally valid measures of mental workload. However, the sensitivity of these measures, as achieved in this study, is low, indicating how the three instructional design conditions, as conceived and implemented, do not impose significantly different mental workload levels on learners.

Keywords: Cognitive Load Theory; Cognitive Load Types; Human Mental Workload; Instructional Design; Direct instructions; Cognitive Theory of Multimedia Learning; Inquiry methods; Community of Inquiry; Reliability; Validity; Sensitivity;

1 Introduction

Cognitive Load Theory (CLT) has been designed to guide instructional designers and practitioners keen to develop instructional resources aimed at promoting the activities of learners, increase their performance and optimise their learning [8,56]. Although CLT has been investigated for many years, developing a set of guidelines aimed at creating effective instructional designs, it has been subjects of multiple critiques due to its theoretical clarity [54] and its methodological approach [15]. In detail, a central problem is the measurement of the overall cognitive load of learners while performing learning activities [43]. Three types of cognitive load have been conceptualised and identified within CLT: intrinsic, extraneous and germane. These are the fundamental assumptions that compose the theory itself. The intrinsic load can be influenced by the familiarity of the learners on a given subject or the intrinsic difficulty and complexity of the learning material to be exploited. The extraneous load can altered by the procedure used to design, organise and deliver instructional material. The germane load is influenced by the effort exerted by learners for handling information, development and automation of schemas in their brains. However, taking into consideration the Popperian's view on critical rationalism [47], CLT cannot be treated a scientific theory due to the lack of clear procedures to measure its fundamental building blocks - the cognitive load types - and thus their empirically validation. As a consequence the theory is believed not to be falsifiable [15]. in other terms, the scientific value of the Cognitive Load Theory and all the other theories built upon the notion of cognitive load [16,15] still lack empirical validation. 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. Unfortunately, although significant advances in Educational Psychology, limited research has been done towards the development of cognitive load assessment techniques [2,1,10,45,6]. The situation is similar in the more specific field of Teaching and Learning [57,19,45].

A domain in which cognitive load has been extensively researched and applied is Ergonomics [64] (Human Factors). In this discipline, the construct of mental workload (MWL), almost overlapping with the construct of cognitive load, has a long history with a plethora of applications for example in the field of aviation [25,24,18] and automotive industry [5]. In these applications, several assessment procedures, both uni-dimensional and multi-dimensional, have been proposed [7,64]. As a consequence, several MWL measures exist in the literature. Similarly, various criteria for validating these measures have been recommended highlighting the continuous interest on MWL research [52]. Taking a broad view, the main logic behind measuring mental workload, in Ergonomics, is to quantify the amount of mental activity devoted to performing a task for predicting human performance and in turn system performance [7]. In Education, the goal is analogous: the logic behind measuring mental workload is to quantify the mental cost of performing a learning task with the objective of predicting the performance of a learner and in turn estimate learning.

This research study is an attempt to bridge the gap present in educational psychology concerned with the measurement of cognitive load by adopting existing measures of mental workload borrowed from Ergonomics. The aim of this study, initiated in [29,42] and extended here, is to evaluate the reliability, validity and sensitivity of three mental workload measures from ergonomics, namely the multidimensional Nasa Task Load Index [18] and Workload Profile [59] as well as the unidimensional Rating Scale Mental Effort [66]. A primary research study has been designed including the comparison of three different instructional design conditions in a third-level master module. The first condition includes the delivery of instructional material using a traditional one-way direct instruction approach (lecturer to students). The second design condition includes the conversion of the instructional material of the first condition into multimedia videos developed by following a set of design principles developed within the Cognitive Theory of Multimedia Learning [36]. The third design condition extends the second design condition by adding a collaborative group activity inspired by the concept of Community of Inquiry [13,4] aimed at extending understanding. Figure 1 summarises this research by presenting its key components, the limitations as emerged in literature, the design of a primary research experiment as well as its evaluation.

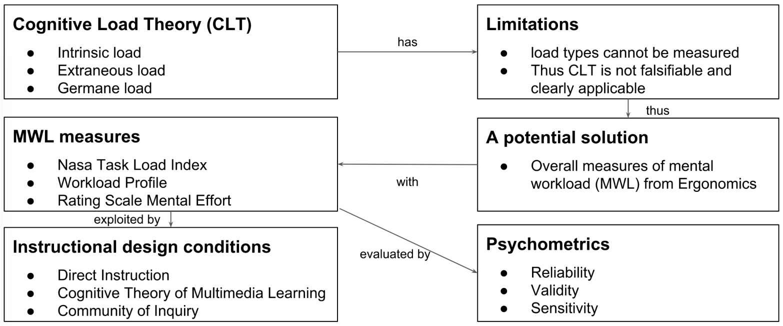

Cognitive Load Theory (CLT)
Intrinsic load
Extraneous load
Germane load
has
Limitations
load types cannot be measured
Thus CLT is not falsifiable and clearly applicable
thus
MWL measures
Nasa Task Load Index
Workload Profile
Rating Scale Mental Effort
with
A potential solution
Overall measures of mental workload (MWL) from Ergonomics
exploited by
Instructional design conditions
Direct Instruction
Cognitive Theory of Multimedia Learning
Community of Inquiry
evaluated by
Psychometrics
Reliability
Validity
Sensitivity

Fig. 1: Summary of the primary research, its motivation and its main components

The rest of the paper is shaped as below. Section 2 introduces Cognitive Load Theory and its cognitive load types. It follows a description of the general issues surrounding other cognitive load-based theories and a brief presentation of state-of-the-art mental workload assessment techniques in Ergonomics, with their advantages and limitations. Afterwards, the paper puts the focus on three self-reporting mental workload procedures because of their adoption in this research. Cognitive Theory of Multimedia Learning, its design principles as well as the Community of Inquiry paradigm, a social constructivist approach for teaching, are described to provide the reader with those relevant notions for understanding the planned experiment. Section 3 focuses on the construction of a primary research experiment with human learners, detailing the methodology and formalising the research hypotheses. Section 4 and 5 respectively present experimental results and critically discuss them emphasising the contribution to the body of knowledge. Section 6 concludes the paper and introduces future work.

2 Theoretical background

2.1 Cognitive load theory

Cognitive Load Theory (CLT) [56] 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 [8]. CLT is an approach that considers the limitations of the information processing system of the human mind [61]. 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 [39] 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 [44]. As a consequence, modelling and assessing cognitive load is far from being a trivial activity. In his seminal contribution, [56] 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 [2,55];
- extraneous load - this is impacted by the way the instructional material is designed, organised and presented [9];
- 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 [44].

Intrinsic cognitive load is considered being static, extraneous load should be minimised [40] and germane load promoted [11]. 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 [10,54,43]. Unfortunately, there is little evidence that these three types are highly separable [12,58,9]. Similarly, to date, there is little evidence about the ways the three different types of load can be coherently and robustly measured [14,43]. According to the traditional critical rationalism proposed by Karl [47], CLT cannot be considered a scientific theory because some of its fundamental assumptions cannot be tested empirically and are thus not falsifiable [15]. 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 [16,15]. 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 [7,61,64].

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 [7,27,49]. 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 [7]. 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 [33,63]. Modern technologies such as web applications have become increasingly complex [23,32,28], with increments in the degree of MWL imposed on operators [17,22]. 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 [7]. 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 [7]. 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.

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 [31,26,41]. Many approaches to operationalise mental workload as a computational concept have been proposed as in [50,41,49,25,30]. Similarly, Various reviews attempted to organise the vast amount of knowledge behind MWL measures and assessment techniques [27,62,7,65]. 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 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 [5,30]. The class of subjective measures include multi-dimensional approaches such as the NASA Task Load Index [18], the Subjective Workload Assessment Technique [48], the Workload Profile [59] as well as uni-dimensional approaches such as the Rating Scale Mental Effort [66], the Subjective Workload Dominance Technique [60] and the Bedford scale [51]. 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 [7,62], 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 [52]. *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 measures

The NASA Task Load Index (NASATLX) instrument is a subjective self-assessment measure of mental workload [18]. It has been extensively applied within Ergonomics in many socio-technical domains, and validated mainly in the transportation industry [18,52]. The measure is built upon six dimensions that are thought to affect mental workload, as described in a number of papers [34,24,29]. Each dimension is assessed with a self-reported judgement by a human, and a weight for each dimension is computed through a paired comparison across dimensions. A subject, after executing a task, is required to express, for each possible pair of the 6 dimensions, (binomial coefficient, $\binom{6}{2} = 15$) , a preference indicating which of the two had a greater contribution to mental workload while executing the underlying task. A weight w for a given dimension is the number of times it was picked as preference in the pairwise procedure. Given the 6 dimensions of the Nasa Task Load Index, each weight is therefore in the range 0 (not relevant) to 5 (more important than any other dimension). The final mental workload score is inferred as a weighed average, taking into account each subjective rating for a dimension d_i and the correspondent weights w_i (equation 1). For comparison purposes in this research, the overall measure is scaled within $[1..100] \in \Re$. The questionnaire can be found in table 13 (appendix).

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

The Workload Profile (WP) is a mental workload assessment procedure [59] developed upon the Multiple Resource Theory [61]. According to this theory, humans are seen as having different capacities or 'attentional 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.

As described in other articles [34,24,29], each dimension is assessed through subjective rates and an individual, after task completion, is required to rate the proportion of attentional resources elicited while performing the task itself. This self-reporting is done expressing a quantity within the range $0..1 \in \Re$. A rating of 0 indicates that the task performed placed no demand while 1 that it required maximum attention. The overall measure of mental worklad is a sum of the 8 rates d (equation 2). For comparison purposes in this research, the overall measure is scaled within $[1..100] \in \Re$. The questionnaire associated to the Workload Profile measure can be found in table 14 (appendix).

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

The Rating Scale Mental Effort (RSME) is a unidimensional mental workload assessment procedure that is built upon the notion of effort exerted by a human over a task. As described in other contributions in the literature [34,24,29], a subjective rating is required by an individual through an indication on a continuous line, within the interval 0 to 150 with ticks each 10 units [66]. Example of labels such as 'absolutely no effort', 'considerable effort' and 'extreme effort' are used along the line (Appendix, table 12). The overall mental workload of an individual coincides to the experienced exerted effort indicated on the line (equation 3. On one hand, although simplicity, the RSME has demonstrated a good degree of sensitivity across different empirical studies. However, on the other hand, it has shown a poor diagnostic power [66].

$$RSME : [0..150] \in \Re \tag{3}$$

2.4 Cognitive Theory of Multimedia Learning

Another cognitivist theory of learning is the Cognitive Theory of Multimedia Learning (CTML) [36,35]. It is strongly connected to other learning theories, including Sweller's Cognitive Load Theory. CTML is based upon three assumptions (figure 2):

- dual-channel assumption - this assumption has been inspired by the dual-coding approach of [46] whereby two separate channels are available for processing information in the human brain, namely the auditory and the visual channel;
- limited processing capacity assumption - in line with the Baddeley's model of working memory [3] and following the assumption of Cognitive Load Theory [56], each channel has a finite, limited capacity;
- active processing assumption - learning is an active process for the selection, filtering, organisation of new information and its integration with prior knowledge.

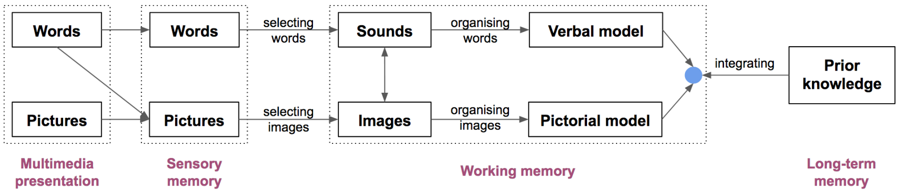


Fig. 2: The model behind Cognitive Theory of Multimedia Learning

Humans are capable of processing a finite amount of information in each channel at a given time. In details, according to CTML, the human brain does not interpret multimedia instructions composed by words, auditory and pictorial information in a mutually exclusive way. Instead, these types of information are firstly selected and then dynamically organised to produce schemas, which are mental logical representations. Schemas are cognitive constructs in which information is organised for storage in long-term memory. Similarly, they can organise simpler elements in a way that these can subsequently act as elements in higher-order schemas. Learning coincides with the development of complex schema as well as the transferring of those learned procedures from controlled processing to automated processing. This shift empties working memory that can then be used for other cognitive processes. [37] suggested five ways for representing words and pictures while information is processed in memory. These are particular stages of processing information. The first stage is represented by words and pictures in the multimedia presentation layer. The second stage includes the acoustic (sounds) and iconic representation (images) in sensory memory. The third stage coincides the sounds and images within working memory. The fourth stage, always within working memory, concerns with the verbal and pictorial models. The fifth stage relates prior knowledge, (the schemas), stored in long-term memory.

Mayer proposed a set of design principles for creating instructions aligned to the above assumptions and stages. Readers can obtain more information on the principles in [38]. Generally speaking, these design principles suggest to provide learners with coherent instructional material in the form of verbal and pictorial information. Coherent information aims to guide learners in the selection of the relevant words and pictures and reduce 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, pre-training, modality; fostering; germane load - multimedia, personalisation, voice, image. These principles have emerged from more than 100 studies conducted in the field [38]. In addition to these, advanced principles have been proposed by Mayer in a number of papers, and recently updated [35]. This demonstrates how CTML is a dynamic theory, suggesting how its principles should not be taken rigidly, but 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 this paper.

2.5 The community of inquiry

A Community of Inquiry (COI) can be defined as a group formed by people interacting within a social context with the goal of investigating the limits of a problematic concept by means of a dialog [13]. 'Dialog' is not a discussion nor a conversation. One one hand, a discussion is a persuasive debate where participants explain their own ideas in at attempt to persuade the other participants. It is a competitive dialectical exchange of ideas that usually ends up with the definition of the correct one, emphasising a winner. On the other hand, a conversation is a spontaneous exchange of ideas and sharing of information. There is no a well-defined way of conversing, leaving learners to develop and build the conversation entirely on their own. The expected outcome is that learners can transfer learned concepts to a new context, and thus expanding their vocabulary and abilities. Instead, a dialog focuses on the thinking of the group as a whole, with the objective of processing certain information aimed both at expanding individual and group knowledge as well as to increase understanding [4].

A pedagogical framework built upon the above definition of dialog is the 'Philosophy for Children' proposed by Mathew Lipman [20] and exploited in the project NORIA [53]. This framework proposes a set of questions aimed at exercising the cognitive abilities of a learner and at developing a higher level of thinking. Lipman, in his work [21], presents a model of reasoning considered to be a genuine and fundamental aspect of any instructional process: the complex thinking. This model is an educational process composed by three ways of thinking: critical, creative and caring thinking. The critical thinking is based upon the formulation of judgements and it is commanded by the criteria of logic, it is self-corrective and sensitive to a context. The dialogue elicits the capacity to think about the thinking (metacognition). In order be understood by others participants within a dialogue, a learner has to clearly explain owns ideas. This communicative requirement leads to a self-correction activity sensitive to the underlying context. The creative thinking is similar to critical thinking in the way of formulating judgements. However, these judgements are strictly related to the underlying context. This type of thinking is self-transcendent and sensitive to the criteria of logic but not governed by them. The caring thinking aims to develop practices regarding the substantial and procedural reflection connected to the resolution of some problem. It is sensitive to the context and it requires metacognitive processes of thinking in order to formulate practical judgments. Within the Community of Inquiry, the development of complex thinking occurs in a process of discovery learning. This process embraces the three type of thinking and it focused on generating and answering philosophical and cognitive questions on logic (critical thinking), aesthetic (creative thinking) and ethic (caring thinking). The Community of Inquiry paradigm has been described for providing the readers with those key notions necessary for the comprehension of the primary research experiment presented in this paper.

3 Design and methodology

A primary research has been designed to investigate the reliability, validity and sensitivity of the three selected subjective mental workload measures (NASA, WP, RSME). 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 taught both to full-time and part-time students. The main difference between full-timers and part-timers is the way classes are planned. 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 and 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 the classroom, have two to three breaks (coffees and lunch). In this research study, conducted over a period of three years (from 2015 to 2017), four topics were delivered to different groups of students, both full-timers and part-timers, in the first part of each academic semester: 'Science', 'The Scientific Method' 'Planning Research' and 'Literature Review'. The remaining topics, taught in the second part of semester, were focused more on practical activities whereby students had to put in practice the theoretical notions provided in the first part of the semester. Three instructional conditions were designed. The first condition included the delivery of instructional material using a traditional one-way direct instructional approach (lecturer to students). The second design condition included the conversion of the instructional material of the first condition into multimedia videos developed by following a set of design principles proposed within the Cognitive Theory of Multimedia Learning [36] (as described in section 2.4). The third design condition extends the second design condition by adding to it a collaborative group activity inspired by the notion of Community of Inquiry [13,4] aimed at extending the understanding of learners (as described in section 2.5). Here the cohort of students is divided into groups composed by 3 or 4 persons performing a collaborative activity. In detail, the differences between the first and the second condition are described in table 15, grouped by the underpinning principles of the CTML. The details of the activity carried out in the third condition are explicated in table 16. Figure 3 and 4 respectively summarise the instructional conditions and the entire research design.

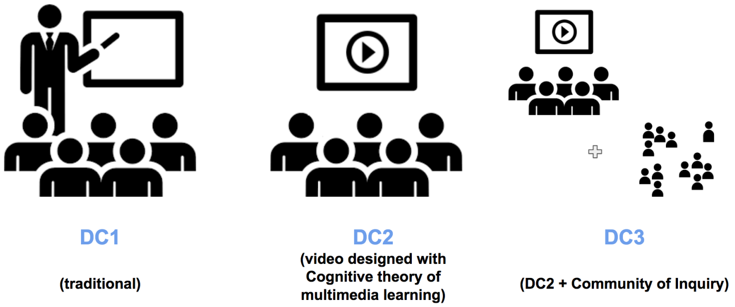

DC1
(traditional)
DC2
(video designed with Cognitive theory of multimedia learning)
DC3
(DC2 + Community of Inquiry)

Fig. 3: Differences between the three instructional design conditions

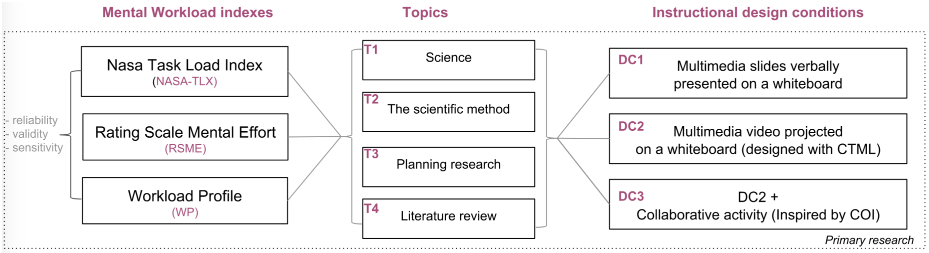

Mental Workload indexes
Topics
Instructional design conditions
reliability
validity
sensitivity
Nasa Task Load Index
(NASA-TLX)
Rating Scale Mental Effort
(RSME)
Workload Profile
(WP)
T1
Science
T2
The scientific method
T3
Planning research
T4
Literature review
DC1
Multimedia slides verbally presented on a whiteboard
DC2
Multimedia video projected on a whiteboard (designed with CTML)
DC3
DC2 + Collaborative activity (Inspired by COI)
Primary research

Fig. 4: Layout of the design of the experiment: three mental workload measures evaluated over three design conditions and three taught topics

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 the case, then the extent to which these measures can discriminate the selected topics, the three instructional conditions as well as the classes delivered will be investigated by computing a measure of their sensitivity. Table 1 lists 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 MWL measures. Therefore these measures would not have reason to exist as participants can 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 mental workload itself, the chances that this occurs are low. Thus, a positive moderate correlation is expected for both types of validity, underlying reasonable relationships between selected MWL measures.

Table 1: Criteria for the evaluation of different mental workload assessment techniques, their definition, associated statistical tests and the expectations for this primary research

Criteria	Definition	Statistical test	expectation
Reliability	the consistency/stability of a MWL measure	Cronbach's Alpha	high
Validity (face)	the extent to which a MWL measure is subjectively viewed as covering MWL itself	Pearson/Spearman correlation	positive & moderate
Validity (convergent)	the degree to which two measures of MWL, theoretically related, are in fact related	Pearson/Spearman correlation	positive & moderate
Sensitivity	the extent to which a MWL measure is able to detect changes in instructional design conditions, topics and classes	ANOVA + t-test/ U-test	moderate

3.1 Participants and procedure

Different cohorts of part-time and full-time students participated in the experimental research and attended the MSc module 'Research design and proposal writing' across different academic semesters between 2015 and 2017. These cohorts of students attended the four topics (T1-T4) listed in figure 4. Some cohort received the first instructional condition (DC1), some other the second (DC2) and some other received the third instructional condition (DC3). At the end of each topic (class), students were asked to fill questionnaires in, aimed at quantifying the mental workload experienced during the class. In details, the three selected self-reporting mental workload assessment techniques, as described in section 2.3, were used in the experimental study: the NASA Task Load Index, the Workload Profile and the Rating Scale Mental Effort. The NASA-TLX and the WP are multi-dimensional and thus require participants to answer a number of questions (figures 13 and 14 in appendix). To facilitate the completion of each questionnaire and not to overwhelm students with many questions, two groups were formed within the same class, one receiving the NASA-TLX and one the WP. Eventually, both the groups received the RSME questionnaire (figure 12 in appendix). The rationale was that, being RSME uni-dimensional, adding one further question to the previous questionnaires was deemed reasonable. In summary, the groups of each class are as below:

- (IRa) MWL instruments received by group A: the NASA-TLX + the RSME
- (IRb) MWL instruments received by group B: the WP + the RSME

Table 2 summarises the number of students across the design condition received, the number of classes for each design condition, across the topics and overall totals.

Table 2: Number of classes, number of students grouped by mental workload instruments received (IRa: NASA-TLX + RSME; IRb: WP + RSME) across design conditions (DC1-3) and topics (T1-4) as well as their totals

	T1			T2			T3			T4			
Design Condition	classes	students		classes	students		classes	#students		classes	students		TOTALS
		IRa	IRb		IRa	IRb		IRa	IRb		IRa	IRb	
DC1	2	13	17	2	20	23	1	11	9	2	20	20	133
DC2	2	23	24	2	16	18	2	22	22	1	13	11	149
DC3	1	9	7	1	10	8	2	15	12	1	9	7	77

TOTALS

Classes	5			5			5			4			19
NASA		45			46			48			42		181
WP			48			49			43			38	178
RSME		45	48		46	49		48	43		42	38	359

The formation of the two groups for each class was random. Groups were planned to be as balanced as possible. However, some of the student who took part in the experimental study did not fully complete the administered questionnaires or they left the class before its administration, therefore associated data was discarded. 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.

4 Results

Table 3 presents the descriptive statistics showing the average (avg), the standard deviation (std) and the Shapiro-Wilk test (W) of normality of the distributions, of the mental workload scores obtained across the different topics and the mental workload assessment techniques (NASA, WP, RSME), grouped by design condition (DC1 - DC3) and topic (T1-T4) along their p-value (p-val). As it is possible to assess from table 3, the p-values (p-val) of the Shapiro-Wilk test (W) obtained for the NASA-TLX and the WP measures are greater than the chosen alpha level ($\alpha = 0.05$), thus, the null hypothesis that the data came from a normally distributed population cannot be rejected (is accepted). However, for the RSME measure, in most of the cases (highlighted), the p-values are lower than the alpha value, thus scores do not follow a normal distribution.

Table 3: Average, standard deviation and Shapiro-Wilk test (W) with p-value (p) at 95% confidence level of the mental workload scores by measure, topic and design condition

		Mental Workload measures								
Topic	Design	NASA			WP			RSME		
		avg	std	W(p)	avg	std	W(p)	avg	std	W(p)
T1	DC1	43.6	08.6	0.96(0.69)	58.6	18.8	0.98(0.99)	42.2	20.5	0.89(0.00)
T2	DC1	51.9	11.9	0.95(0.40)	51.9	15.1	0.95(0.27)	57.0	23.0	0.97(0.25)
T3	DC1	50.2	12.8	0.91(0.25)	50.2	15.9	0.91(0.29)	54.9	20.8	0.90(0.04)
T4	DC1	48.3	11.5	0.95(0.32)	56.6	10.9	0.96(0.61)	53.3	20.8	0.97(0.25)
T1	DC2	41.8	17.2	0.98(0.90)	49.2	15.2	0.95(0.34)	45.4	18.6	0.95(0.03)
T2	DC2	50.2	10.8	0.97(0.86)	57.2	09.7	0.95(0.36)	62.0	17.3	0.94(0.06)
T3	DC2	43.5	12.2	0.96(0.43)	51.9	14.1	0.94(0.20)	46.5	18.2	0.94(0.02)
T4	DC2	52.2	16.4	0.96(0.74)	45.5	19.2	0.90(0.17)	59.0	19.0	0.91(0.04)
T1	DC3	38.5	11.3	0.94(0.54)	60.0	14.8	0.85(0.12)	38.0	22.6	0.94(0.31)
T2	DC3	48.2	12.5	0.97(0.88)	50.8	10.3	0.89(0.25)	65.1	24.2	0.85(0.01)
T3	DC3	45.4	14.6	0.98(0.93)	58.6	13.1	0.96(0.77)	51.8	20.2	0.92(0.04)
T4	DC3	48.3	07.6	0.98(0.96)	60.3	11.7	0.96(0.78)	60.3	24.5	0.85(0.01)

4.1 Reliability

To assess the reliability of the selected mental workload measures, the Cronbach's Alpha has been employed. It measures the internal consistency of the items of a multidimensional 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-dimensional. Table 4 shows the Cronbach's Alpha coefficients of the other two selected multidimensional mental workload measures (NASA-TLX and the WP), across all the topics (T1-T4) and the instructional design conditions (DC1-DC3). 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-TLX and the WP can be considered reliable respectively with a coefficient of 0.73 and 0.847. To confirm this high reliability, Cronbach's Alpha has been computed also across each topic and instructional condition (table 5). The alpha scores are mostly above 0.6 for the NASA-TLX and 0.8 for the WP strongly suggesting how these measures have an inherent good reliability.

Table 4: Overall reliability of the multidimensional mental workload measures with sample size, related number of items in the scales and associated Cronbach's Alpha

Instrument	Sample size	# of items	Cronbach's α
NASA	181	6	0.730
WP	178	8	0.847

Table 5: Reliability of the multidimensional MWL measures computed with the Cronbach's α, grouped by topic (T1-4) and design condition (DC1-3)

Topic	Design	Mental Workload measures			
		NASA-TLX		WP	
	condition	Size	α	Size	α
T1	DC1	13	0.63	17	0.91
T2	DC1	20	0.69	23	0.87
T3	DC1	11	0.59	9	0.93
T4	DC1	20	0.65	20	0.81
T1	DC2	23	0.84	24	0.83
T2	DC2	16	0.56	18	0.67
T3	DC2	22	0.66	22	0.81
T4	DC2	13	0.81	11	0.92
T1	DC3	9	0.72	7	0.88
T2	DC3	10	0.79	8	0.64
T3	DC3	15	0.80	12	0.83
T4	DC3	9	0.24	7	0.80

4.2 Validity

To assess the validity of the three MWL measures, two sub-forms have been selected, namely face and convergent validity. The former validates the extent to which a MWL measures is subjectively viewed as covering the construct of MWL itself while the latter validates the degree to which two measures of MWL, expected to be theoretically related, are in fact related. To assess face validity, a question on overall MWL has been designed and asked to students straight after the completion of each class and before starting to fill the MWL questionnaires in (figure 17). The answers to this new question have been correlated to the scores of the selected MWL measures (NASA-TLX, WP, RSME), as listed in table 6.

Table 6: Face validity of the mental workload assessment instruments, namely the Nasa Task Load Index, The Workload Profile and the Rating Scale Mental Effort, the sample size, the Pearson and Spearman correlation coefficients

Instrument	Sample size	Pearson r	Spearman ρ
NASA	181	0.49	0.47
WP	178	0.39	0.40
RSME	359	0.42	0.41

To assess convergent validity, the MWL scores produced by the multidimensional NASA-TLX and the WP measures have been correlated against the MWL scores of the unidimensional RSME measure. This test was possible because a participant filled in either the questionnaire associated to the NASA-TLX or WP, and at the same time the RSME. Correlation between the NASA-TLX and WP cannot be computed because no participant received the questionnaires associated to these measures at the same time. Both the Pearson (parametric) and the Spearman's Rank (non-parametric) correlation coefficients have been employed for computing validity. Both parametric and non-parametric tests have been employed because not all the distributions of table 3 were normal. Tables 6, 7 respectively shows the correlations for face validity and convergent validity.

Table 7: 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	181	0.49	0.47
WP vs RSME	178	0.29	0.31

4.3 Sensitivity

The sensitivity of the selected MWL measures has been computed by checking whether the distributions of their scores are statistically significant different across the topics (T1-T4), the instructional design conditions (DC1-DC3) and the classes (C1-19). Figure 5 depicts the boxplots of these distributions for visual inspection.

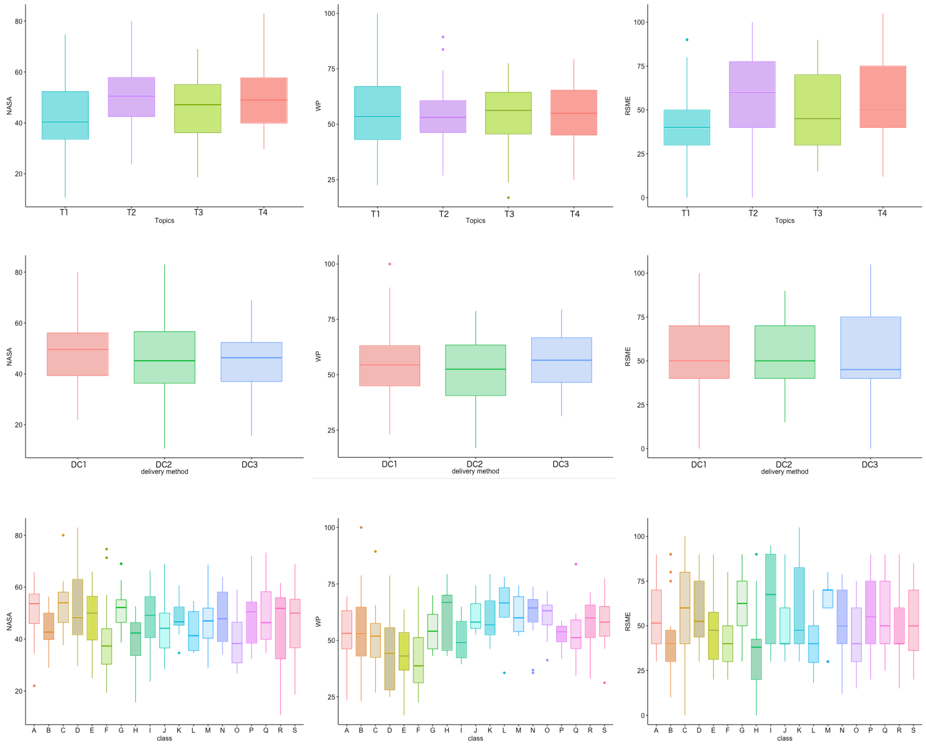


Fig. 5: Boxplots of the distributions of the mental workload scores by measure (NASA, WP, RSME) grouped by topic (T1-T4), design condition (DC1-3) and class (A-S)

Formally, a Kruskal-Walllis analysis with a 95% confidence interval has been conducted. This is equivalent to a one-way analysis of variance on ranks and it is a non-parametric method for testing whether samples originate from the same distribution. This has been chosen because not all the distributions of table 3 are normal. As it is possible to see from table 8, some statistical significant difference was spotted across topic and classes, but not for instructional design conditions.

Table 8: Comparison of distributions of the workload scores using the Kruskal-Wallis test with 95% confidence interval (Chi-squared, degrees of freedom and p-values)

Group by	NASA			WP			RSME		
	X^2	DF	p-val	X^2	DF	p-val	X^2	DF	p-val
topic (T1-T4)	10.91	3	0.012	0.22	3	0.973	35.66	3	<0.0001
design condition (DC1-3)	2.44	2	0.293	3.43	2	0.179	0.146	2	0.9290
class (C1-19)	20.25	18	0.318	33.30	18	0.015	45.42	18	0.0003

The Kruskal-Walllis test does not precisely tells which distributions are statistically significantly different. Thus, the Wilcoxon-Matt-Whitney test (or Mann-Whitney U-test) was employed only where a difference was spotted by the Kruskal-Wallis test. It is a non-parametric test for comparing the means of two groups that are not normally distributed. Table 9 lists the comparisons across topics of the NASA-TLX and RSME scores. Tables 10 and 11 respectively list the comparisons of the WP and RSME scores by classes. From table 9, the NASA-TLX was able to produce scores significantly different twice across six comparisons while the RSME five times out of six, demonstrating higher sensitivity across topics. The WP, out of all the possible comparisons across classes, was able to produce scores significantly different 22 times out of 171 (table 10), while the RSME 46 out of 171 (table 11), showing a higher sensitivity across classes.

Table 9: P-values of the pairwise U-test with 95% confidence interval by topic

Topic	NASA			RSME		
	T1	T2	T3	T1	T2	T3
T2	0.002	-	-	<0.00001	-	-
T3	0.196	0.095	-	0.020	0.0006	-
T4	0.014	0.596	0.241	<0.0001	0.237	0.0241

Table 10: P-values of the pairwise U-test with 95% confidence interval by class for the Workload Profile scores

Class	WP																	
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
B	0.93	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
C	1.00	0.684	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
D	0.47	0.27	0.36	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
E	0.29	0.13	0.38	1.00	-	-	-	-	-	-	-	-	-	-	-	-	-	-
F	0.28	0.06	0.19	0.93	0.70	-	-	-	-	-	-	-	-	-	-	-	-	-
G	0.72	0.75	0.38	0.13	0.03	0.02	-	-	-	-	-	-	-	-	-	-	-	-
H	0.36	0.43	0.20	0.12	0.02	0.03	0.58	-	-	-	-	-	-	-	-	-	-	-
I	0.88	0.56	1.00	0.38	0.23	0.09	0.35	0.16	-	-	-	-	-	-	-	-	-	-
J	0.37	0.53	0.23	0.24	0.08	0.06	0.31	0.81	0.18	-	-	-	-	-	-	-	-	-
K	0.26	0.29	0.18	0.06	0.02	0.02	0.37	0.94	0.11	0.83	-	-	-	-	-	-	-	-
L	0.11	0.39	0.06	0.09	0.03	0.02	0.12	0.72	0.13	0.71	0.53	-	-	-	-	-	-	-
M	0.15	0.31	0.12	0.09	0.01	0.01	0.18	0.72	0.09	1.00	0.83	0.58	-	-	-	-	-	-
N	0.10	0.34	0.06	0.07	0.01	0.01	0.18	0.78	0.12	1.00	0.86	0.72	0.96	-	-	-	-	-
O	0.15	0.27	0.06	0.01	0.01	0.01	0.20	0.75	0.07	1.00	0.89	0.39	1.00	0.64	-	-	-	-
P	1.00	1.00	0.86	0.20	0.14	0.06	0.61	0.34	0.63	0.28	0.16	0.03	0.14	0.03	0.02	-	-	
Q	0.96	0.83	0.80	0.17	0.19	0.09	0.66	0.52	1.00	0.37	0.26	0.15	0.14	0.09	0.06	0.93	-	
R	0.38	0.51	0.17	0.10	0.02	0.01	0.55	0.65	0.21	0.69	0.82	0.22	0.48	0.36	0.57	0.18	0.27	-
S	0.35	0.64	0.25	0.12	0.03	0.02	0.49	0.91	0.22	0.64	0.79	0.34	0.67	0.54	0.84	0.33	0.41	0.93

Table 11: P-values of the pairwise U-test with 95% confidence interval by class for the Rating Scale Mental Effort scores

Class	RSME																	
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
B	0.09	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
C	0.43	0.03	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
D	0.46	0.01	0.81	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
E	0.27	0.33	0.07	0.04	-	-	-	-	-	-	-	-	-	-	-	-	-	-
F	0.05	0.95	0.01	0.01	0.34	-	-	-	-	-	-	-	-	-	-	-	-	-
G	0.25	0.01	0.86	0.58	0.01	0.01	-	-	-	-	-	-	-	-	-	-	-	-
H	0.02	0.49	0.01	0.01	0.10	0.33	0.01	-	-	-	-	-	-	-	-	-	-	-
I	0.19	0.01	0.59	0.42	0.02	0.01	0.60	0.01	-	-	-	-	-	-	-	-	-	-
J	0.50	0.45	0.24	0.14	1.00	0.45	0.09	0.18	0.11	-	-	-	-	-	-	-	-	-
K	0.57	0.03	0.91	0.84	0.15	0.02	0.62	0.01	0.58	0.20	-	-	-	-	-	-	-	-
L	0.03	0.88	0.01	0.01	0.17	0.57	0.01	0.67	0.01	0.36	0.03	-	-	-	-	-	-	-
M	0.23	0.06	0.75	0.57	0.03	0.01	0.95	0.01	0.52	0.20	0.96	0.01	-	-	-	-	-	-
N	0.62	0.24	0.18	0.16	0.60	0.22	0.06	0.07	0.05	0.69	0.31	0.10	0.14	-	-	-	-	-
O	0.14	0.76	0.04	0.01	0.60	0.77	0.01	0.26	0.01	0.60	0.04	0.41	0.02	0.29	-	-	-	-
P	0.67	0.07	0.71	0.83	0.13	0.02	0.58	0.01	0.27	0.35	0.80	0.01	0.59	0.28	0.06	-	-	-
Q	0.91	0.10	0.35	0.35	0.32	0.06	0.17	0.02	0.13	0.45	0.56	0.03	0.30	0.57	0.12	0.57	-	-
R	0.38	0.28	0.08	0.05	0.93	0.37	0.01	0.09	0.03	0.88	0.15	0.25	0.02	0.55	0.57	0.16	0.32	-
S	0.80	0.21	0.33	0.28	0.46	0.14	0.14	0.06	0.09	0.67	0.35	0.07	0.30	0.78	0.22	0.50	0.82	0.46

5 Discussion

Two multidimensional and a unidimensional subjective mental workload (MWL) measures, 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 [18] and the Workload Profile [59] while the latter is the Rating Scale Mental Effort [66]. These measures have been applied in a typical third-level classroom in the context of a module taught in the School of Computing, at the Dublin Institute of Technology. The experiment included the quantification and analysis of the experienced mental workload of different cohorts of students who were exposed to three different instructional design conditions and four topics. An analysis of the reliability of the two multidimensional MWL measures has been performed through a quantification of their internal consistency. In details, Cronbach's Alpha has been employed to assess the relation of the items associated to each MWL assessment technique. An obtained alpha value of 0.73 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 0.847. 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, extended also to the selected unidimensional MWL measure, namely the Rating Scale Mental Effort. 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 (NASA-TLX), the Workload Profile (WP) and the Rating Scale Mental Effort (RSME) - are subjectively viewed as covering the construct of MWL itself by students. The latter validity indicates the degree to which the two multidimensional measures of MWL are theoretically related with the unidimensional measure. The obtained positive Pearson and Spearman correlation coefficients suggest how the three MWL measures are moderately correlated to the overall mental workload self-reported by students (correlations between $0 - 39 - 0.49$), thus demonstrating, as expected, moderate face validity. Similarly, the achieved positive Pearson and Spearman correlation coefficients show the expected moderate relationships that exist between the two multidimensional MWL measures (NASA-TLX and WP) and the unidimensional MWL measure (RSME), thus demonstrating moderate convergent validity. Eventually, with highly reliable and moderately valid MWL measures, their sensitivity was subsequently computed. Sensitivity referred to the extent to which the three selected MWL measures were able to detect changes of MWL scores across the four topics, the three instructional design conditions and the nineteen classes delivered over a period of 3 years. In detail, sensitivity was assessed through a non-parametric analysis of the variance of the MWL scores by adopting the Kruskal-Walllis test. This test was able to detect some statistical significant difference of the MWL scores across topic and classes, but not for instructional design conditions. Subsequently, an extended analysis of these detected differences was performed by a pairwise comparison of the MWL distributions employing the Wilcoxon-Matt-Whitney test (or Mann-Whitney U-test). The test showed how the NASA Task Load index was able to detect some of the differences in MWL scores only across topics while the Workload Profile only across classes. However, the unidimensional Rating Scale Mental Effort scale succeeded in detecting differences in MWL scores across topics and classes. None of the three measures was able to detect differences in MWL scores across the design conditions suggesting how they did not really impacted the variation of mental workload experienced by students. Figure 6 summarises the findings visually comparing the reliability, validity and sensitivity of the three selected mental workload measures.

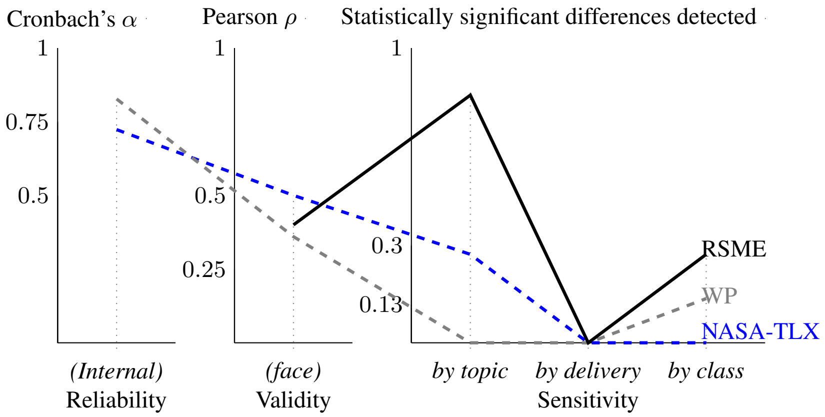


Fig. 6: Comparison of the reliability, validity and sensitivity of the Nasa Task Load Index (NASA-TLX), the Workload Profile (WP) and Rating Scale Mental Effort (RSME)

Intuitively, given the strong reliability and moderate validity achieved by these measures, it is reasonable to infer that the design principles from the Cognitive Theory of Multimedia Learning - applied to design the second instructional condition - and the application of the Community of Inquiry approach - employed to design the third instructional condition - were, in this primary research, as not effective as expected, despite the different expectation. This research contributes to the body of knowledge by offering an alternative application of existing measures of mental workload, mainly adopted within Ergonomics, in Education, and in particular within the field of Teaching and Learning. Additionally, the experiment proposed in this study is in line to the Popperian's view of falsifiability because it is transparent and can be replicated and eventually falsified. Every attempt aimed at falsifying the findings achieved in this research is not seen as a negative pursuit but rather a positive endeavour because it is aimed at increasing our understanding of mental workload as a construct applied within Education, Teaching and Learning for evaluating the efficiency of various instructional approaches.

6 Conclusions

The research conducted in this paper was an attempt to investigate the reliability, validity and sensitivity of three well known self-reporting mental workload (MWL) measures, mainly used within Ergonomics, within third-level education. A primary research study has been designed and executed to gather self-reported data by different cohort of students of a post-graduate module in Computer Science. In details, four different topics of a module on 'research design and proposal writing' were repeatedly delivered using three different instructional approaches over a period of 3 years. The first design approach included the delivery of theoretical material by employing a traditional direct instruction method employing slides projected to a white-board that included textual and pictorial information. The second approach included the delivery of the same theoretical material through multimedia videos built by employing a set of principles from Cognitive Theory of Multimedia Learning [38]. The third design approach included the extension of the second approach with a collaborative group activity for students inspired by the Community of Inquiry paradigm [21]. Evidence strongly suggests how the three MWL measures are reliable when applied in a typical third-level classroom. Results demonstrated their moderate validity, in line with the validity achieved in other empirical experiments within Ergonomics. On the contrary, their sensitivity was very low in discriminating the mental workload scores of the three different instructional design conditions. However, given the high reliability and modest validity of the three MWL measures, the achieved sensitivity might be reasonably attributed to the minimal impact of the way the three instructional design conditions were designed.

Future work will include the replication of this primary research across other instructional design conditions, topics and third-level modules as well as the development of a hybrid scale that takes into account the strengths and limitations of the three mental workload assessment instruments adopted in this research.

References

1. Artino Jr, A.R.: Cognitive load theory and the role of learner experience: An abbreviated review for educational practitioners. Aace Journal 16(4), 425–439 (2008)
2. Ayres, P.: Using subjective measures to detect variations of intrinsic cognitive load within problems. Learning and Instruction 16(5), 389–400 (2006)
3. Baddeley, A., Hitch, G.: Working memory, vol. 8, pp. 47–90. Academic Press (1974)
4. Bleazby, J.: Autonomy, democratic community, and citizenship in philosophy for children: Dewey and philosophy for children's rejection of the individual/community dualism. Analytic teaching 26(1), 30–52 (2006)
5. Brookhuis, K.A., de Waard, D.: Monitoring drivers' mental workload in driving simulators using physiological measures. Accident Analysis & Prevention 42(3), 898–903 (2010)
6. Brunken, R., Plass, J.L., Leutner, D.: Direct measurement of cognitive load in multimedia learning. Educational psychologist 38(1), 53–61 (2003)
7. Cain, B.: A review of the mental workload literature. Tech. rep., Defence Research and Development Canada Toronto (2007)
8. Chandler, P., Sweller, J.: Cognitive load theory and the format of instruction. Cognition and Instruction 8(4), 293–332 (1991)
9. 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? Computers in Human Behavior 25(2), 315–324 (2009)
10. De Jong, T.: Cognitive load theory, educational research, and instructional design: some food for thought. Instructional science 38(2), 105–134 (2010)
11. Debue, N., van de Leemput, C.: What does germane load mean? an empirical contribution to the cognitive load theory. Frontiers in Psychology 5, 1099 (2014)
12. DeLeeuw, K.E., Mayer, R.E.: 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 (2008)
13. Dewey, J.: The child and the curriculum. No. 5, University of Chicago Press (1902)
14. Dixon, P.: From research to theory to practice: Commentary on chandler and sweller. Cognition and Instruction 8(4), 343–350 (1991)
15. Gerjets, P., Scheiter, K., Cierniak, G.: The scientific value of cognitive load theory: A research agenda based on the structuralist view of theories. Educational Psychology Review 21(1), 43–54 (2009)
16. Goldman, S.R.: On the derivation of instructional applications from cognitive theories: Commentary on chandler and sweller. Cognition and Instruction 8(4), 333–342 (1991)
17. Gwizdka, J.: Distribution of cognitive load in web search. Journal of the american society & information science & technology 61(11), 2167–2187 (November 2010)
18. Hart, S.G.: Nasa-task load index (nasa-tlx); 20 years later. In: Human Factors and Ergonomics Society Annual Meeting. vol. 50, pp. 904–908. Sage Journals, San Francisco, California, USA (2006)
19. Kirschner, P.A.: Cognitive load theory: Implications of cognitive load theory on the design of learning. Learning and instruction 12(1), 1–10 (2002)
20. Lipman, M., Sharp, A.M., Oscanyan, F.S.: Philosophy in the classroom. Philadelphia: Temple University Press (1980)
21. Lipman, M.: Thinking in education. Cambridge University Press (2003)
22. Longo, L.: Human-computer interaction and human mental workload: Assessing cognitive engagement in the world wide web. In: INTERACT (4). pp. 402–405 (2011)
23. Longo, L.: Formalising human mental workload as non-monotonic concept for adaptive and personalised web-design. In: Masthoff, J., Mobasher, B., Desmarais, M., Nkambou, R. (eds.)

User Modeling, Adaptation, and Personalization. Lecture Notes in Computer Science, vol. 7379, pp. 369–373. Springer (2012)

24. Longo, L.: Formalising Human Mental Workload as a Defeasible Computational Concept. Ph.D. thesis, Trinity College Dublin (2014)
25. Longo, L.: A defeasible reasoning framework for human mental workload representation and assessment. Behaviour and Information Technology 34(8), 758–786 (2015)
26. Longo, L.: Designing medical interactive systems via assessment of human mental workload. In: Int. Symposium on Computer-Based Medical Systems. pp. 364–365 (2015)
27. 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 (June 2016)
28. Longo, L.: Subjective usability, mental workload assessments and their impact on objective human performance. In: IFIP Conference on Human-Computer Interaction. pp. 202–223. Springer (2017)
29. Longo, L.: 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. In: Proceedings of the 10th International Conference on Computer Supported Education, CSEDU 2018, Funchal, Madeira, Portugal, March 15-17, 2018, Volume 2. pp. 166–178 (2018)
30. Longo, L., Barrett, S.: Cognitive effort for multi-agent systems. In: International Conference on Brain Informatics, Toronto, Canada. Lecture Notes in Computer Science, vol. LNCS 6334, pp. 55–66. Springer (2010)
31. Longo, L., Barrett, S.: A computational analysis of cognitive effort. In: Intelligent Information and Database Systems, Second International Conference, ACIIDS, Hue City, Vietnam. Lecture Notes in Computer Science, vol. LNCS 5991, pp. 65–74. Springer (2010)
32. Longo, L., Dondio, P.: 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. pp. 345–352 (2015)
33. Longo, L., Kane, B., Hederman, L.: Argumentation theory in health care. In: 25th International Symposium on Computer-Based Medical Systems, Rome, Italy. pp. 1–6. IEEE (2012)
34. Longo, L., Rusconi, F., Noce, L., Barrett, S.: The importance of human mental workload in web-design. In: 8th International Conference on Web Information Systems and Technologies, Porto, Portugal. pp. 403–409. SciTePress (April 2012)
35. Mayer, R.: Using multimedia for e-learning. Journal of Computer Assisted Learning 33(5), 403–423 (2017), jCAL-16-266.R1
36. Mayer, R.E.: Multimedia learning. Psychology of Learning and Motivation 41, 85–139 (2002)
37. Mayer, R.E.: The Cambridge handbook of multimedia learning. Cambridge university press (2005)
38. Mayer, R.E.: Multimedia learning. Cambridge University Press (2009)
39. Miller, G.A.: The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological Review 63(2), 81–97 (1956)
40. Mousavi, S., Low, R., Sweller, J.: Reducing cognitive load by mixing auditory and visual presentation modes. Journal of Educational Psychology 87(2), 319–334 (1995)
41. Moustafa, K., Luz, S., Longo, L.: Assessment of mental workload: a comparison of machine learning methods and subjective assessment techniques. In: Int. Symposium on Human Mental Workload: Models and Applications. pp. 30–50 (2017)
42. Orru, G., Gobbo, F., O'Sullivan, D., Longo, L.: 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, March 15-17, 2018, Volume 2. pp. 292–302 (2018)
43. Paas, F., Tuovinen, J.E., Tabbers, H., Van Gerven, P.W.: Cognitive load measurement as a means to advance cognitive load theory. Educational psychologist 38(1), 63–71 (2003)
44. Paas, F., Van Merrienboer, J.J.G.: 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 (1993)
45. Paas, F.G., Van Merriënboer, J.J., Adam, J.J.: Measurement of cognitive load in instructional research. Perceptual and motor skills 79(1), 419–430 (1994)
46. Paivio, A.: Mental Representations: A Dual Coding Approach. Oxford Psychology Series, Oxford University Press (1990)
47. Popper, K.: Conjectures and refutations: The growth of scientific knowledge. routledge (2014)
48. Reid, G.B., Nygren, T.E.: The subjective workload assessment technique: A scaling procedure for measuring mental workload. In: Hancock, P.A., Meshkati, N. (eds.) Human Mental Workload, Advances in Psychology, vol. 52, chap. 8, pp. 185–218. North-Holland (1988)
49. 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, pp. 215–229. Springer International Publishing, Cham (2016)
50. 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: 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), Bari, Italy, November 16-17, 2017. pp. 126–140 (2017)
51. Roscoe, A.H., Ellis, G.A.: A subjective rating scale for assessing pilot workload in flight: A decade of practical use. Technical report TR 90019, Royal Aerospace Establishment (March 1990)
52. Rubio, S., Diaz, E., Martin, J., Puente, J.M.: Evaluation of subjective mental workload: A comparison of swat, nasa-tlx, and workload profile methods. Applied Psychology 53(1), 61–86 (2004)
53. Satiro, A.: Jugar a pensar con mitos: este libro forma parte del Proyecto Noria y acompaña al libro para niños de 8-9 anos: Juanita y los mitos. Octaedro (2006)
54. Schnotz, W., Kürschner, C.: A reconsideration of cognitive load theory. Educational Psychology Review 19(4), 469–508 (2007)
55. Seufert, T., Jänen, I., Brünken, R.: The impact of intrinsic cognitive load on the effectiveness of graphical help for coherence formation. Computers in Human Behavior 23(3), 1055–1071 (2007)
56. Sweller, J., Van Merrienboer, J., Paas, F.: Cognitive architecture and instructional design. Educational Psychology Review 10(3), 251–296 (1998)
57. Sweller, J.: Cognitive load theory, learning difficulty, and instructional design. Learning and instruction 4(4), 295–312 (1994)
58. Sweller, J.: Element interactivity and intrinsic, extraneous, and germane cognitive load. Educational psychology review 22(2), 123–138 (2010)
59. Tsang, P.S., Velazquez, V.L.: Diagnosticity and multidimensional subjective workload ratings. Ergonomics 39(3), 358–381 (1996)
60. Vidulich, M.A., Ward Frederic G., S.J.: Using the subjective workload dominance (sword) technique for projective workload assessment. Human Factors Society 33(6), 677–691 (December 1991)
61. Wickens, C.D.: Multiple resources and mental workload. Human Factors 50(2), 449–454 (2008)

62. Wilson, G.F., Eggemeier, T.F.: Mental workload measurement. In: Karwowski, W. (ed.) Int. Encyclopedia of Ergonomics and Human Factors (2nd ed.), vol. 1, chap. 167. Taylor and Francis (2006)

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

64. 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)

65. Young, M.S., Stanton, N.A.: Mental workload: theory, measurement, and application. In: Karwowski, W. (ed.) Encyclopedia of ergonomics and human factors, vol. 1, pp. 818–821. Taylor & Francis, 2nd edn. (2006)

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

Appendix

Table 12: The Rating Scale Mental Effort

Please indicate, by marking the horizontal axis below, how much effort it took for you to execute the task you have just completed.

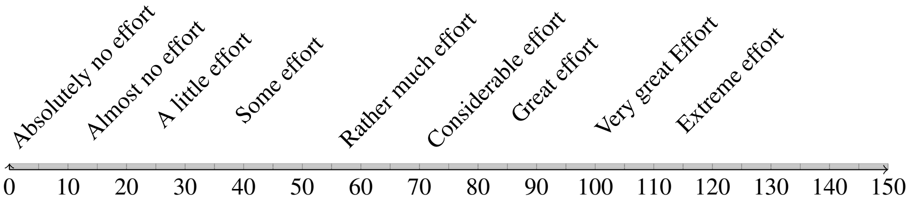


Table 13: The NASA Task Load Index (NASA-TLX)

Label	**Question**
NT_1	How much mental and perceptual activity was required (e.g. thinking, deciding, calculating, remembering, looking, searching, etc.)? Was the task easy or demanding, simple or complex, exacting or forgiving?
NT_2	How much physical activity was required (e.g. pushing, pulling, turning, controlling, activating, etc.)? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?
NT_3	How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic?
NT_4	How hard did you have to work (mentally & physically) to accomplish your level of performance?
NT_5	How successful do you think you were in accomplishing the goals, of the task set by the experimenter (or yourself)? How satisfied were you with your performance in accomplishing these goals?
NT_6	How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task?

Table 14: The Workload Profile (WP)

Label	**Question**
WP_1	How much attention was required for activities like remembering, problem-solving, decision-making, perceiving (detecting, recognising, identifying objects)?
WP_2	How much attention was required for selecting the proper response channel (manual - keyboard/mouse, or speech - voice) and its execution?
WP_3	How much attention was required for spatial processing (spatially pay attention around)?
WP_4	How much attention was required for verbal material (eg. reading, processing linguistic material, listening to verbal conversations)?
WP_5	How much attention was required for executing the task based on the information visually received (eyes)?
WP_6	How much attention was required for executing the task based on the information auditorily received?
WP_7	How much attention was required for manually respond to the task (eg. keyboard/mouse)?
WP_8	How much attention was required for producing the speech response (eg. engaging in a conversation, talking, answering questions)?

Table 15: Design of the instructional condition 2 using the principles of Cognitive Theory of Multimedia Learning and its differences with condition 1 grouped by load type

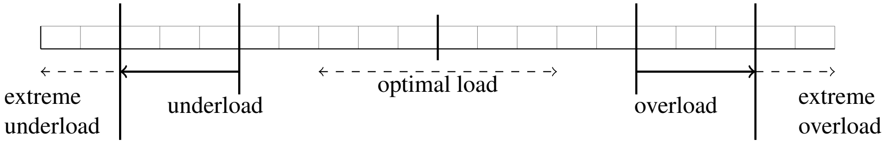
Principle	Load type	Design condition 1	Design condition 2
coherence	extraneous	any extraneous material was kept to minimum.	
signaling	extraneous	cues, in the form of relevant keywords, with a larger font size	cues (relevant keywords), popped-in in the video to emphasise the organisation of essential 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/pictures. Other times, the image was removed and important sentences were textually presented full screen.

Table 16: Dialogical activity set for the third design condition inspired by the Community Inquiry paradigm

Which are the most important concepts explained during the lesson? Through a dialogue with the members of your team, talk about these concepts, try to define them and try to eliminate misunderstandings

Table 17: Question and scale designed for investigating the face validity of the mental workload assessment measures

How much mental workload the teaching session imposed on you?



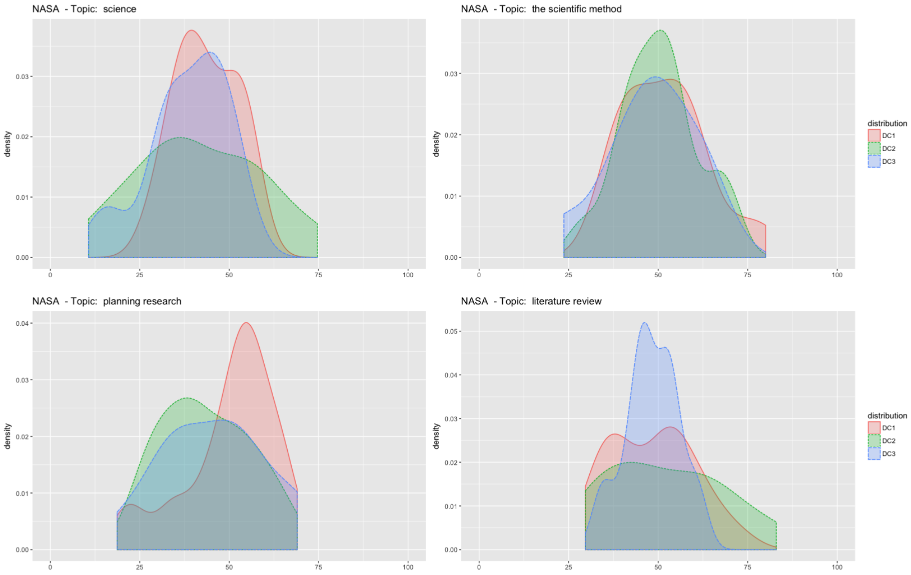


Fig. 7: Density plots of the distributions of the mental workload scores by topic (T1-T4) and design condition (DC1-3) for the NASA Task Load Index

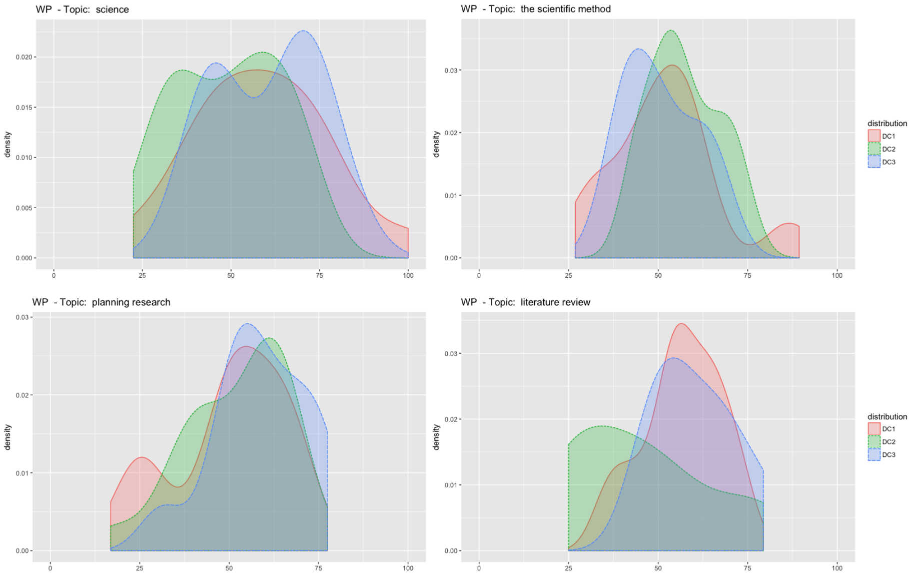


Fig. 8: Density plots of the distributions of the mental workload scores by topic (T1-T4) and design condition (DC1-3) for the WorkloadProfile

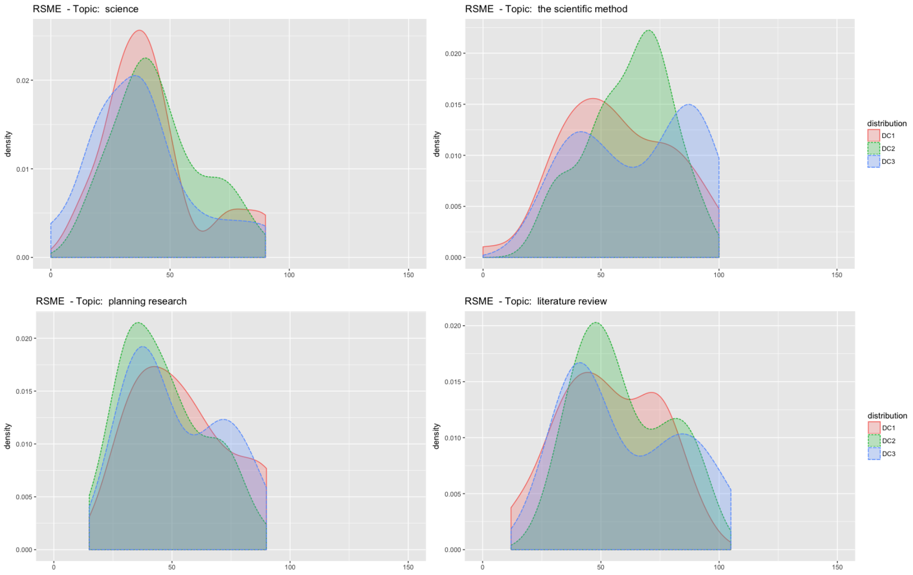


Fig. 9: Density plots of the distributions of the mental workload scores by topic (T1-T4) and design condition (DC1-3) for the Rating Scale Mental Effort

View publication stats