

2020-12-23

Evaluating instructional designs with mental workload assessments in university classrooms

Luca Longo

Technological University Dublin, luca.longo@tudublin.ie

Giuliano Orru'

Technological University Dublin

Follow this and additional works at: <https://arrow.tudublin.ie/scschcomart>



Part of the [Applied Behavior Analysis Commons](#)

Recommended Citation

Luca Longo & Giuliano Orrú (2020) Evaluating instructional designs with mental workload assessments in university classrooms, *Behaviour & Information Technology*, pages = {1-31}, DOI: 10.1080/0144929X.2020.1864019

This Article 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 Articles 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](#)

Evaluating instructional designs with mental workload assessments in university classrooms

Luca Longo^a, Giuliano Orrú

School of Computer Science, Technological University Dublin, Republic of Ireland

ARTICLE INFO

Keywords:

Cognitive Load Theory
Instructional Design
Cognitive Theory of Multimedia Learning
Subjective Mental Workload.

ABSTRACT


Cognitive Load Theory (CLT) has been conceived for improving instructional design practices. Although researched for many years, one open problem is a clear definition of its cognitive load types and their aggregation towards an index of overall cognitive load. In Ergonomics the situation is different with plenty of research devoted to the development of robust constructs of mental workload (MWL). By drawing a parallel between CLT and MWL, as well as by integrating relevant theories and measurement techniques from these two fields, this paper is aimed at investigating the reliability, validity and sensitivity of three existing self-reporting mental workload measures when applied to long learning sessions, namely the NASA Task Load index, the Workload Profile and the Rating Scale Mental Effort, in a typical university classroom. These measures were aimed at serving for the evaluation of two instructional conditions. Evidence suggests these selected measures are reliable and their moderate validity is in line with results obtained within Ergonomics. Additionally, an analysis of their sensitivity by employing the descriptive Harrell-Davis estimator suggests that the Workload Profile is more sensitive than the Nasa Task Load Index and the Rating Scale Mental Effort for long learning sessions.

1. Introduction

Cognitive Load Theory (CLT), a cognitivist theory, has been initially conceived as a form of guidance for instructional designers [68] eager to develop instructional material that is presented in a manner that promotes the activities of learners and optimise their performance as well as their learning [7]. CLT is a wider framework that takes into account the limitations of the information processing system of the human mind [78]. Intuitively, the assumption is that if a learner is either underloaded or overloaded, learning is likely to be hampered (figure 1). The main assumption of CLT is that the human cognitive architecture is limited in its capacity and its main function to process and store information is finite and it has a direct consequence on learning [43]. Also, the experience of cognitive load is highly subjective, different from human to human, and influenced by the learner's education and training as well as own cognitive style [53]. As a consequence, modelling and assessing cognitive load is far from being a trivial activity [33, 5]. In his seminal contribution, [68] have initially proposed three types of cognitive load. The intrinsic load is influenced by the unfamiliarity of the learners or the intrinsic complexity of the learning material under use [1, 67]. The extraneous load is impacted by the way the instructional material is designed, organised and presented [8]. The germane load is influenced by the effort exerted to deal and to process information, to construct and automate schemas in the brain of the learners [53]. It is believed that germane load should be promoted, the extraneous load should be minimised and that the intrinsic load is static and cannot be changed for a given learning task [45, 11]. These three types of load have gone through three decades of evolution and redefini-

tion [51]. After a number of critiques related to the theoretical development of CLT and after several failed attempts to develop generally applicable measure of the three types of load, CLT has been re-conceptualised using the notion of element interactivity: the amount of elements that have to be simultaneously processed in working memory. In detail, elements of the task should be learned while task unrelated elements should be discarded for a successful schema construction. With this new notion, the definition of the intrinsic and extraneous loads were updated [70]. The extraneous load is nowadays considered to be the level of interactivity of the elements composing an instructional material used for teaching activities. Instructional designs should be aligned to it and should not focus on enhancing the number of items to be processed by learners, otherwise the resulting load could be considered extraneous [51]. In other words, if instructional designs do not include instructions that increase the number of elements that have to be processed by learners within their working memory, more spare capacity exist for supporting learning. In this specific case, existing instructions can facilitate the use of working memory that is allocated for the intrinsic load. Eventually, the germane load is no longer an independent source of load. Rather, it depends on those working memory resources related to the intrinsic load of a learning task. As a consequence, intrinsic load depends on the characteristic of a learning task, extraneous load is influenced by the characteristics of the instructional material as well by the characteristic of the instructional design and on the prior knowledge of learners. Germane load depends on the characteristics of a learner that influence the allocation of resources of working memory devoted to the intrinsic load [70] (figure 2).

Cognitive Load Theory, although highly relevant for instructional design and with a plethora of applications in the last decades, providing a series of effects and guidelines

 luca.longo@tudublin.ie (L. Longo)
ORCID(s): 0000-0002-2718-5426 (L. Longo)

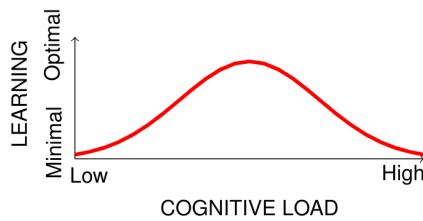


Figure 1: The theoretical connection between human performance (learning) and overall cognitive load

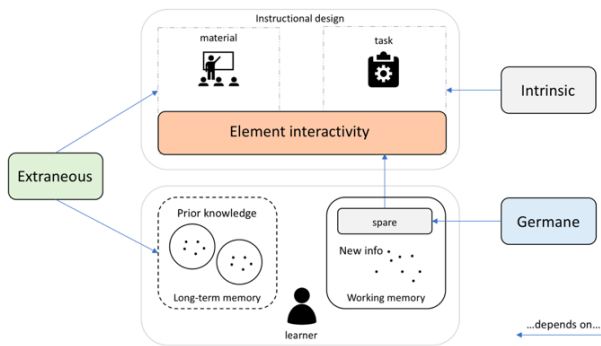


Figure 2: Redefinition of the cognitive load types and their roles [51].

to create efficient instructional designs, it has been criticised because of its theoretical clarity [66] and its methodological approach [17]. In detail, a fundamental, open and challenging problem exists: the measurement of its three cognitive load types [51, 10, 66, 52] and the measurement of the cognitive load of learners during learning tasks [52]. There is little evidence that these three types are highly separable [12, 70, 8] and minimal knowledge exist on the ways they can be consistently measured and aggregated [14, 52]. Because of this, and by considering the critical rationalism brought forward by Popper [55, 56], some authors believe CLT cannot be deemed scientific because its fundamental assumptions, its cognitive load types, cannot be directly tested empirically and therefore cannot be falsified [17]. In other words, it lacks empirical validation and therefore cannot be considered scientific. However, others believe that the cognitive load types can actually be tested empirically by manipulating instructional conditions experimentally. This rationale is also due because subjective measures of cognitive load types are problematic because they depend on learner perception, therefore it is better to depend on experimental manipulation. Uncertainty also exists on the methods for aggregating these types towards an overall index of cognitive load. Therefore, as acknowledged by other authors, the main research challenge concerns the development of valid and reliable measure of overall cognitive load, to empirically demonstrate the scientific value of CLT as well as all the other theories built upon it [18, 52, 17]. On one hand, CLT and related measures of its types have been mainly brought forward by educational psychologists and evolved for more than three decades

[93, 51]. On the other hand, within the discipline of Ergonomics and Human Factors more effort has been dedicated to the development of overall cognitive load assessment techniques. In this discipline, cognitive load is mainly referred to as human Mental Workload (MWL), a well known psychological construct [89] with several several applications in the aviation [24, 22] and automotive industries [4]. In these domains, many measurement techniques, both uni-dimensional and multi-dimensional, have been developed for MWL assessment [92, 6, 89]. Similarly, various criteria for validating these techniques have been proposed during the last five decades [65]. In Ergonomics, the principal reason for measuring and assessing mental workload, is to quantify the mental cost associated to executing a task with the goal of predicting operator/learner and system performance [6]. In Education the situation is similar: the main reason for measuring cognitive load is to quantify the mental cost exerted to perform a learning task in order to predict the learner's performance and thus trying to assess learning. By drawing a parallel between CLT employed within education and MWL within ergonomics, and by integrating relevant theories and measurement techniques from these two fields of research, this paper is aimed at investigating the reliability, validity and sensitivity of existing overall MWL assessment techniques, when applied to long learning sessions.

A primary research is designed aimed at comparing two instructional designs in a third-level post-graduate course by means of mental workload measurement. The first condition includes the delivery of instructional material by employing the traditional teacher-directed instruction method where slides are projected to a white-board to learners. The second condition is a 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 (CTML) [37, 38]. CLT and CTML are greatly connected by various underlying assumptions including the dual-modality information processing channels, the limited working memory capacity and the development of schemata in long-term memory. However, it has to be noted that the former theory mainly focuses on cognitive load and its impact on learners to acquire new knowledge. Whereas the latter focuses on the role of different types of cognitive processing while learning and not on the working memory load generated by these processes [30, 41]. For these reasons, the construct of mental workload, borrowed from ergonomics has been employed relaxing this distinction. Here, mental workload is assessed as overall load imposed by information processing. Three mental workload measures have been selected: the multidimensional Nasa Task Load Index [24] and Workload Profile [73] and the unidimensional Rating Scale Mental Effort [94]. The investigated research question is: *to what degree can self-reporting measures of mental workload, from Ergonomics, contribute to the evaluation of instructional designs based*

on the traditional verbal direct-instruction approach and the multimedia learning approach?

The reminded of this paper continues in Section 2 by reviewing state of the art measures of mental workload within Ergonomics with a discussion on their advantages and limitations. A detailed description of the selected self-reporting mental workload assessment techniques follows, along with the Cognitive Theory of Multimedia Learning and its design principles. Section 3 provides further details on the design of the primary research experiment involving human learners, detailing the methodology and research hypotheses. Section 4 present the results of the experiment followed by a critical discussion. Section 5 concludes this paper highlighting the contribution to the body of knowledge and setting future directions for research.

2. Theoretical background

The construct of mental workload has a long history in the fields of psychology and ergonomics with several applications in the aviation [24, 22] and automotive industries [4]. Although it has been studied for the last five decades, no clear definition of MWL has emerged that has a general validity and that is universally accepted [6] The principal reason for measuring and assessing MWL is to quantify the mental cost of performing a certain task with the goal of predicting operator and system performance [6]. A measure of MWL can be considered as an extremely powerful design criterion. In fact, at the initial stage of design, not only can a system/interface be developed with mental workload into consideration, but a measure of MWL can also guide designers in designing and implementing relevant structural variations [88]. For example, in modern technologies, such as web applications, becoming increasingly complex, there is an increment in the degree of MWL imposed on operators [19, 20], thus design alternatives have to be evaluated at different design phases [34]. The main belief usually considered in design approaches is that as the task complexity and difficulty increases, MWL also should increase and performance decrease [6]. Consecutively, errors have a higher frequency, response times are longer, and fewer tasks are completed within an established unit of time [27]. However, when task difficulty is minimal, systems might impose a very low mental workload on operators. This case should also be avoided as it likely leads to complications in preserving attention and increase reaction time [6]. Modeling the construct of Mental Workload is not easy. Many approaches to aggregate factors believed to influence mental workload have been proposed, both deductive as in [59, 61, 60], and inductive as in [47, 46]. Regardless of the aggregation strategies three main classes of measures, as described in the following sections, exist: self-reporting, task-performance measures and physiological measures. The field of research devoted to the development of measures of mental workload is large, scattered and extremely challenging 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 Several reviews attempted to organise the vast amount of knowledge related to measurement procedures. In general, the measurement techniques which have emerged in the literature can be clustered into three large categories [79, 90, 72, 74, 86, 6, 91]:

- *self-assessment, subjective measures*: it includes self-reporting rating scales or questionnaires.
- *task objective performance measures*: divided into primary and secondary task measures, they mainly refers to objective indicators of task performance;
- *physiological measures*: they are derived from human physiology and responses of the body.

The class of *self-report measures*, often referred to as subjective measures, relies on the subjective perceived experience of the interaction operator-system. They are referred to as subjective because the focus of such measures may be subjective, that means a perception of a specific factor by an individual. They have always appealed many workload practitioners because it is strongly believed that only the person concerned with a task can provide an accurate and precise judgement with respect to the mental workload exerted. This category of measures comprise multi-dimensional approaches for mental workload measurement such as the NASA Task Load index [24], the Workload Profile [73], the Subjective Workload Assessment Technique [58] as well as uni-dimensional approaches such as the Subjective Workload Dominance Technique [77], the Instantaneous Self-Assessment of workload [71], the Bedford scale [62] and the Rating Scale Mental Effort [94]. Various dimensions and factors influencing mental workload are accounted for as for instance, task-related as mental and temporal demands, person-specific individual characteristics including the emotional state, motivation and general attitude of an operator [4]. These measures usually include close-ended scales and, when case multidimensional, they comprise an aggregation strategy that combines the different operator answers into an overall scalar of mental workload. The class of *task performance measures* is of utility for those practitioners and designers mainly interested in the performance of their systems and interactive technologies. In this context, the belief is that the mental workload exerted by a human while interacting with a certain interface, technology or a complex system, becomes essential only if it influences system performance. Therefore, it is believed that this class is the most valuable option for user-centred designers [74]. Performance measures are primary and secondary. In the former type, the performance of an operator is monitored and evaluated according to variations in primary-task demands. Whereas in the latter type, the performance of a human on the secondary task might not have practical

relevancy, but it can serve to load the primary task or to measure the mental workload of the person executing it [6, 79, 90, 86]. 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 evaluating their effect on the body's states, and not by measuring performance on the primary tasks or self-reported ratings. Examples include heart rate, pupil dilation and blinking, blood pressure, brain activation and muscle signals as measured respectively by electroencephalograms and electromyograms [87].

Self-reporting measures are in most of the times of easy administration and analysis. They usually provide designers with an overall score of mental workload. Additionally, multi-dimensional measures can help designer trace back the main sources of mental workload. These measures are usually administered after the execution of a task, however, they influence the reliability in the case of long tasks. Sometimes they are administered during task execution but the drawback is that they might influence the execution itself, becoming a source of workload. Additionally, meta-cognitive constraints can reduce the precision of reporting making it arduous to perform comparisons on an absolute scale among raters. However, from the literature, it seems they are the most appropriate measures for assessing mental workload because they have demonstrated high levels of sensitivity and diagnosticity [65]. *Task performance measures* can be primary or secondary. Primary-task measures represent a direct indicator of human performance and they are usually precise for long tasks. They are capable of discerning individual differences because of different cognitive resource allocation strategies. However, the main drawback relates to their poor capacity in distinguishing human performance on multiple parallel tasks. If they are taken individually, they have demonstrated poor reliability, but if used with other classes of measures, they have been proven useful. Secondary task measures are usually good in discriminating tasks when primary performance measures cannot help find differences. This is mainly due to the fact that they are useful for quantifying the spare attentional capacity of an individual for short tasks. Unfortunately, they are usually intrusive and sensitive to large changes in mental workload, influencing the behaviour of an operator while performing the primary task. *Physiological measures* are suitable for monitoring continuous signals. They have demonstrated high sensitivity and they tend not to interfere with the human performance on the primary task. The main limitation is that they can be influenced by external artifacts and interference and they usually required equipment that is often physically obtrusive. Eventually, the analysis of these measures is very complex, requiring the presence of well trained experts and engineers for setting the equipment and interpret the gathered signals. However, with advances in sensor-based technology, these limitations have been becoming weaker as less obtrusive and more

precise equipment is available and accessible to designers and people without expertise or an engineering background.

In this study, the class of measures that have been considered is the self-reporting class. The rationale behind this decision is that these measures are easy to be administered at the end of a typical university classroom. Primary task measures, such as multiple choice questionnaires, can indeed be helpful as they represent a direct indication of classroom learning, but unfortunately they require the intervention of a lecturer for setting them up as they are topic specific. Secondary task measures are believed to affect the behaviour of learners in a typical university classroom. Eventually, physiological measures require equipment to be attached to the body or scalp of each learner, demanding significant setting time that is not available in a typical university class. The next sections are devoted to the detailed description of three mental workload assessment techniques, their formalism to produce a score of mental workload.

2.1. Assessing Mental workload via self-reporting measures

Many self-reporting mental workload assessment techniques exist. In this section, the three that have been selected for experimental purposes are further described. The NASA Task Load Index is a type of self-assessment subjective measure [22]. It has been validated in the aviation industry and in other contexts in Ergonomics [22, 65] with several applications in many socio-technical domains. It is a combo of six factors that can highly influence mental workload (questions of table 10). Each of this is self-reported with a close-ended subjective judgement and a paired comparison among factors determines the weight (importance) for each judgement. Learners are required to choose, 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 'physical or mental Demand?', 'frustration or performance?' and so forth. The weights w coincides with how many times each dimension has been selected. Since there are 15 comparisons, the weight of a factor is in the range 0 (not relevant) to 5 (more important than any other factor). The overall mental workload index is a weighed average of the self-reporting ratings, one for each attribute d_i multiplied by the correspondent weight w_i :

$$NASA TLX = \left(\sum_{i=1}^6 d_i \times w_i \right) \frac{1}{15}, \quad [0..100] \in \mathfrak{R}$$

An alternate version exist in the literature that do not consider the pair-wise comparison procedure and its derived weights: the RAW-TLX [23, 48].

$$RAW TLX = \sum_{i=1}^6 d_i, \quad [0..100] \in \mathfrak{R}$$

It has been shown that a high correlation between the weighted and unweighted workload indexes can be often

obtained thus some author suggest to use the simpler raw version [44, 13]. In this study, we preferred to use the original version due to its potential for providing more diagnostic information for analysis.

The Workload Profile (WP) mental workload assessment procedure [73] is built upon the Multiple Resource Theory proposed in [78, 79]. Here, individuals as learners have different cognitive capacities or 'resources'. These include the *stage of information processing* which can be perceptual/central and/or related to the response selection/execution; the *code of information processing* which can be spatial and/or verbal; the *input modality* refers to visual and/or auditory processing; the *output modality* can be either manual and/or verbal (speech). Each resource is quantifiable via a self-reported scale (questions of table 11) and humans, after each task completion, are requested to self-assess and quantify the proportion of attentional resources employed for the execution of a task (with a value in the range $0..1 \in \mathfrak{R}$). A rating of 0 is meant to refer to that situation in which a task placed no demand whatsoever on the human, while 1 expresses that the task required maximum attention on that resource. The final score is due by summing the 8 rates d (averaged here, and scaled in $[1..100] \in \mathfrak{R}$ for comparison purposes):

$$WP : [0..100] \in \mathfrak{R} \quad WP = \frac{1}{8} \sum_{i=1}^8 d_i \times 100$$

The uni-dimensional Rating Scale Mental Effort (RSME) is measure that assesses effort with the assumption that it is highly correlated to mental workload. This scale take into consideration for the exerted effort by a human during the execution of a task. It can be reported across a continuous scale (intervals within the range 0 to 150 and ticks each 10 units, Appendix 8). Labels such as 'a little effort' and 'rather much effort' are used along the continuous scale. The final mental workload is the exerted effort self-reported by a learner, from the origin of the scale (zero).Formally:

$$RSME : [0..150] \in \mathfrak{R}$$

On one hand, although the scale is simple and fast to be administered, it has demonstrated good sensitivity [94]. However, on the other hand, it has shown a poor diagnostic power [94]. For further details about the scale, its history, and development, the reader is referred to [94].

2.1.1. Assessment in educational contexts

Research exists at the intersection of mental workload measurement applied in educational contexts. For example, Wiebe et al. [80] examined the NASA-TLX and the Subjective Cognitive Load measure [52] and assessed the relative efficacy in the design of multimedia-based educational environments. Findings firstly showed how the weighted version of the NASA-TLX had minimal additional value when compared to its unweighted counterpart. Secondly, both the measures were sensitive in both extraneous and intrinsic loads, and they showed differences. Authors suggested

that an account and better understanding of germane load is essential to improve the mental workload utility in instructional design.

2.2. Cognitive Theory of Multimedia Learning

A popular cognitivist theory of learning is the Cognitive Theory of Multimedia Learning (CTML). It has been conceived and developed by Prof. Mayer [37, 38, 36]. This theory is strictly supported by other learning theories such as the Cognitive Load Theory [69]. CTML is based upon three assumptions. I) dual-channel assumption: two separate channels can be used for processing information in the brain, namely the auditory and the visual channel, in line with the dual-coding approach of [54]. II) the assumption of limited processing capacity: each channel is limited in its capacity, this be aligned also with the assumption of CLT [68] and the working memory model of Baddeley [2]. III) active processing assumption: learning is considered to be an active process involving the selection of information, its filtering and organisation as well as its integration with prior knowledge. Humans, in each channel, can process a finite set of pieces of information at a time. According to the CTML, multimedia instructions either made by words, audios or pictures are not interpreted by the brain independently and mutually exclusively. Rather, these individual representations of information are selected and subsequently dynamically organised to produce coherent mental logical representations called *schemas*.

These are particular cognitive constructs able to organise information for storage in long-term memory. In detail, 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 subsequently employed for other mental processes. Mayer suggested five ways of representing words and pictures while information is processed in memory [39]. These are particular stages of processing information (as depicted in figure 3). The first is the pictures and words in the layer of multimedia presentation. 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 model of verbal and pictorial information, 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 design principles (table 1) for creating instructions aligned to the above limitations of the brain and the dual-channel paradigm of learning. These principles are aimed at supporting the design of coherent instructional material for learners as a combination 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

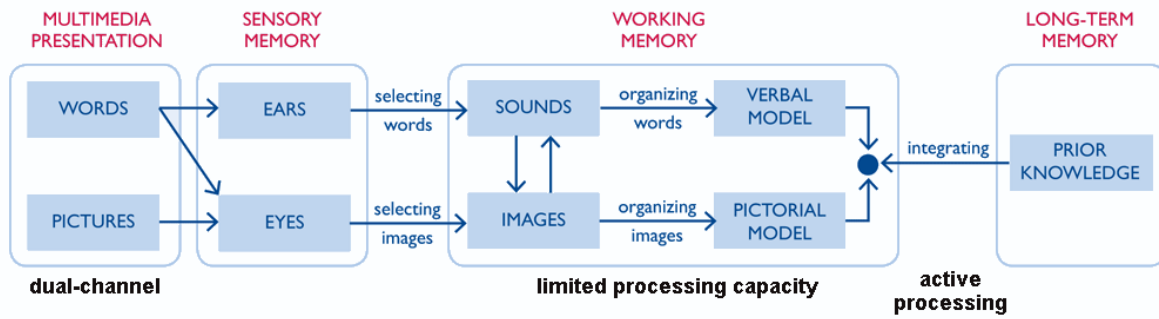


Figure 3: The model developed within the Cognitive Theory of Multimedia Learning [42]

strictly connected to the Cognitive Load Theory because its twelve principles can be linked to the original three types of loads. In fact, it has been suggested that the principles that minimise extraneous load are: coherence, signaling, redundancy, spatial contiguity, temporal contiguity; Those that manage intrinsic load are: segmenting, pre-training, modality fostering; and those that optimise germane load are: multimedia, personalisation, voice, image (as per table 1). Note that since there is uncertainty surrounding the three cognitive load types, which are now believed to be two [30, 28], as described in the introduction, then this categorisation should be taken mainly as a form of guidance. These principles have emerged from more than hundred of empirical research studies [40] and have evolved into more advanced principles [36]. This evolution exhibits the dynamism surrounding this theory and it suggests how the principles should not be taken rigidly, but rather as a starting point for discussion and experimentation as in this study. CTML has been mainly described to provide readers with those key elements necessary for the comprehension of the primary research experiment presented in the next section.

3. Research design, methods and hypotheses

A primary research experiment has been designed to investigate the reliability and the validity of the three aforementioned self-reporting mental workload assessment techniques (NASATLX, WP, RSME) as well as their sensitivity to discriminate different design conditions. Experiments have been conducted in the School of Computer Science at the (XXX university - Blind review), 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 part-time 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. At the begin-

ning of each semester, four topics were presented to learners: 'Science', 'The Scientific Method' 'Planning Research' and 'Literature Review'. The rationale behind the selection of these topics 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 students, by employing direct instructions methods. In other words, this would have facilitated the application of the three selected self-reporting mental workload assessment techniques at the end of the delivery of each topic, without interruptions or unexpected events. 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 classes. In fact, part-time students, given the long classes, were expected to loose interest during the day, with a constant reduction of their engagement and the effort exerted towards learning. In contrast, full-timers were expected to better maintain attention, given the 2-hour classes they were exposed to. 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 rationales, two design conditions were eventually formed. These conditions were built according to the design principles behind the Cognitive Theory of Multimedia Learning (CTML) - as described in section 2.2. In detail, the differences between the two design conditions are described in table 2, grouped by the underpinning principles of the CTML (table 1). Figure 4 synthesises and depicts the full research design.

Table 1

Design Principles of Cognitive Theory of Multimedia Learning and their theoretical relation to the load types of Cognitive Load Theory [40] according to earlier conceptualisation with three types of load

Principle	Description	Reference to load type (CLT)
Coherence	humans better learn when instructions do not contain extraneous material	extraneous
Signaling	learning is increased when explicit cues, for highlighting the organisation of the essential instruction, are added to the instructional material	extraneous
Redundancy	humans' learning is better promoted by only using narration and graphical aids than by using narration jointly with graphical aids and printed text	extraneous
Spatial Contiguity	learning is greatly improved when corresponding pictures and words are closely placed in space and not in separate screen locations or pages	extraneous
Temporal Contiguity	humans better learn when corresponding pictures and words are presented simultaneously rather than presented at different stages over time	extraneous
Segmenting	learning is increased when multimedia-based instructions are delivered in user-paced segments and not as a single continuous unit	intrinsic
Pre-training	humans learning is greatly enhanced from multimedia instructions when a pre-training is offered providing to learners with key names and characteristics of the instructional components	intrinsic
Modality	learning is enhanced when graphical aids and verbal narration are employed rather than graphics and printed textual information	intrinsic
Multimedia	humans learn better from pictures and words than from only words	germane
Personalisation	learning is increased by a multimedia presentation when words are presented in a conversational manner rather than using a formal style	germane
Voice	humans learning is enhanced from multimedia-based instructions when words are narrated by a friendly human voice rather than by an artificial machine	germane
Image	learning is not necessarily improved by multimedia instructions only because the speaker's image is displayed on the screen rather than when is not	germane

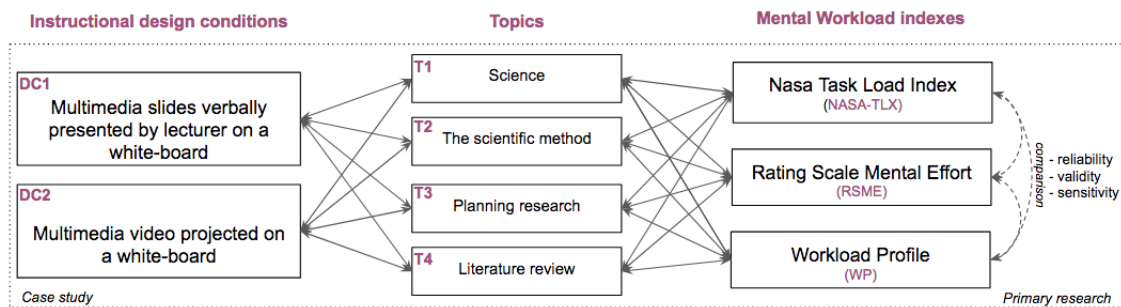


Figure 4: Outline of and key components of the empirical research experiment

3.1. Research hypotheses

Informally, the research hypotheses are that the NASA Task Load Index (NASATLX), the Workload Profile (WP) and the Rating Scale Mental Effort (RSME) 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 instruments can discriminate the two design conditions will be investigated by computing a measure of their sensitivity. Table 3 shows the research hypotheses, lists the criteria for evaluating different mental workload mea-

asures, 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 could precisely assess mental workload autonomously. Similarly, a high degree of convergent validity would imply that two different measures assess the construct of mental workload

Table 2

The design conditions and their differences in terms of adherence to the principles of Cognitive Theory of Multimedia Learning

Principle	Screenshots	Design condition (A)	Design condition (B)
coherence	fig. 17	extraneous instructional material was maintained to a minimum level.	
signaling	fig. 18	cues, as relevant keywords in the text, were made more visible by using a larger font size	cues, as relevant keywords in the narration, appeared in the video in order to emphasise the organisation of the essential material.
redundancy	fig. 19	use of graphical aids and the lecturer did not fully read the content of the slides but narrated them	the majority of words of the printed text was eliminated in order to offload the visual channel (eyes); use of graphical aids from lecturer who narrated the content.
spatial contiguity	fig. 20	corresponding words and pictures were placed close to each other in space and not in different slides.	
temporal contiguity	fig. 21	corresponding pictures and words were presented at the same stage over time	corresponding pictures and words (verbally transmitted) were presented at the same stage over time.
segmenting	fig. 22	the instructions are presented as a single unit	the instructions are presented in different segments and they are separated by visual transition effects.
pre-training	-	pre-training was not offered to learners.	
modality	fig. 23	printed text is kept in the slides and verbally explained	printed text is eliminated in order to offload the visual channel (eyes) and verbally explained loading another channel (ears)
multimedia	fig. 24	words and pictures.	
personalisation	fig. 25	words are presented by using a friendly conversational tone and not by using a formal style	
voice	-	words are verbally pronounced by a human lecturer and not by an artificial tool	
image	fig. 26	no video has been employed, therefore no speaker's image was available. However, the lecturer was in the classroom, thus present.	the lecturer's image was most of the time kept in the video, sometimes using the full space available, sometimes 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 full screen.

Table 3

Criteria for the evaluation of different mental workload (MWL) assessment techniques, their definition, the associated statistical tests and the expectations for this primary research

Criteria	Description	Correspondent statistical test	expectation
Reliability	the stability or consistency of a MWL measure	Cronbach's Alpha	high
Validity (face)	the extent to which a MWL measure is subjectively viewed as covering the construct of MWL itself	Pearson/Spearman correlation	positive and moderate
Validity (convergent)	the degree to which two measures of MWL, expected to be theoretically related, are in fact related	Pearson/Spearman correlation	positive and moderate
Sensitivity	the extent to which a MWL measure allows the detection of changes in instructional design conditions	ANOVA + T-test/Wilcoxon test	moderate

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 different mental workload measures.

3.2. Procedure and participant demographics

Distinct groups of full-time and part-time post-graduate learners participated in the experimental study and attended

the post-graduate module 'Research design and proposal writing' in various semesters. In detail, learners attended the four topics listed in figure 4 (T1-T4), not necessarily in the same order, depending on the semester. Classes were delivered either using the first instructional condition (DC1) or the second (DC2). Thus, all learners in each class were exposed to one and only one condition, for a given topic. After the delivery of each topic, and with no distinction of design condition, students were invited to compile question-

naires to obtain information necessary to quantify their mental workload experienced while attending the class. In detail, three self-reporting mental workload measures (as in section 2.1) were employed:

- the NASA Task Load Index (NASA-TLX, questionnaire in table 10);
- the Workload Profile (WP, questionnaire in table 11);
- the Rating Scale Mental Effort (RSME, questionnaire in figure 8).

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 receiving the WP questionnaire. Eventually, both the groups received the RSME questionnaire. The rationale was that, being RSME uni-dimensional, adding one further question to the other two questionnaires was deemed reasonable. In synthesis, the two subgroups are:

- sub-group A (NASA-TLX + RSME);
- sub-group B (WP + RSME).

A study information sheet was presented to students as well as a consent form to sign. This documentation was approved by the ethics committee of the (XXX University, Blind review). 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 during a class. Table 4 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 or not completed at all. This was because some of them left the classroom before its end, or because of incorrect filling of the questionnaire. In the latter case, if most of the questions were unanswered, the observation was discarded. However, if only a small percentage of questions were unanswered, the observation was kept and empty answers were imputed from the rest of the class. In detail, on one hand, for the NASA-TLX, some data was missing for the pair-wise comparison procedure, thus logistic regression was used for imputing the binary preference. On the other hand, concerning the WP instrument, linear regression was used for imputing omitted answers. Also note that, on one hand, the length of the classes for the four topics delivered with design condition 1 (table 4) slightly differ. This is because the lecturer did not read the slides but narrated them at different peaces in different days and class-rooms. Also, full control on the context in which the class took place, and the people, was not possible. In fact, some minor inconvenient occurred such as rebooting the computer or the projector, some attempt from students to ask questions or to repeat. However, except technical problems with technology, and to guarantee

Table 4

Description of topics, design conditions, sub-groups, mental workload instruments received and number of students for each sub-group, and each class, with length in minutes. For example, line 1 refers to the topic 'Science', delivered in design condition 1 twice: the first delivery had 7 students who received the NASA+RSME and 11 the WP+RSME, and it lasted 62 mins; the second had 6 students who received the NASA+RSME while 6 the WP+RSME, and lasted 60 mins.

Topic	(# of students)		Class length (mins)
	A (NASA + RSME)	B (WP + RSME)	
Design condition DC1:			
T1 - Science	7/6	11/6	62/60
T2 - The scientific method	10/11	13/10	46/46
T3 - Planning research	11/11	9/10	54/21
T4 - Literature Review	10/11	11/10	55/33
Totals	77	80	377
Design condition DC2:			
T1 - Science	13/10/13	13/11/9	17
T2 - The scientific method	12/6	12/6	28
T3 - Planning research	11/11	11/11	10
T4 - Literature Review	13/9	11/9	18
Totals	98	93	163
Overall totals DC1+DC2	175	173	540

that the intrinsic load was maintained as similar as possible across classes, no fragments of content was repeated and no questions were answered during the delivery of the content. An opportunity to have feedback to students was only addressed after the end of each experimental task, when all experimental data was collected. On the other hand, the execution time for design condition 2 was always the same, as these are multimedia pre-recorded videos. Someone can argue that time might influence cognitive load. A recent study attempted to investigate the impact of processing time on cognitive load [57]. However, in this study, the time dimension has not been taken into consideration. The rationale is that this hypothesis has been mainly tested for short tasks [57], leading to uncertain findings while the learning tasks under consideration are long. Also, according to [57], one variable that likely affect processing time, and thus cognitive load, is prior knowledge, not available in the present study. Overall, 540 minutes of class delivery was performed (9 hours) by the same lecturer (377 for DC1, 163 for DC2), over 6 semesters (3 academic years), across the four topics.

4. Results

Table 5 presents the descriptive statistics of each sub-group introduced in table 4. In detail, it shows the average (avg), the standard deviation (std), the median (med) and the Shapiro-Wilk test (W) of normality of the distribu-

Table 5

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	Mental Workload assessment technique														
	NASA					WP					RSME				
	#	avg	SD	med	W/p	#	avg	SD	med	W/p	#	avg	SD	med	W/p
Design condition 1															
science	13	43.62	8.64	42.67	0.96/0.69	17	58.6	18.81	60	0.98/0.99	30	42.2	20.58	40	0.89/0
scientific method	21	52.29	11.55	52.67	0.96/0.48	23	51.83	15.11	51.88	0.94/0.22	44	57.36	22.87	56.5	0.97/0.3
planning research	22	46.27	13.77	51.17	0.91/0.04	19	51.31	18.02	54.38	0.89/0.03	41	52.51	21.97	50	0.94/0.03
literature review	21	47.12	12.46	49.67	0.97/0.65	21	56.49	10.75	56.25	0.97/0.67	42	53.87	19.47	50.54	0.97/0.28

Design condition 2

science	36	43.12	15.2	42.83	0.98/0.85	33	47.99	17.28	50	0.95/0.14	69	44.06	18.07	40	0.95/0.01
scientific method	18	47.69	12.65	48.17	0.96/0.69	18	57.29	9.79	55.94	0.95/0.36	36	61.67	17.93	67.5	0.93/0.02
planning research	22	43.55	12.23	42.33	0.96/0.43	22	51.93	14.16	56.56	0.94/0.2	44	46.59	18.26	40	0.94/0.02
literature review	22	50.82	15.05	51.33	0.97/0.7	20	48.88	20.24	46.88	0.98/0.86	42	58.38	21.63	57.5	0.95/0.07

tions, along the p-values (p-val) of the mental workload scores obtained across the different topics and the mental workload techniques (NASA, WP, RSME), grouped by design condition (DC1, DC2) and topic (T1-T4). Looking at table 5, the majority of the p-values (p-val) as well as the Shapiro-Wilk test scores (W) are greater than the chosen alpha level ($\alpha = 0.05$). Therefore, for most of the subgroups, the hypothesis of data coming from a normally distributed population cannot be rejected. This means that the mental workload scores associated to most of the topics, follow a normal distribution. Figures 10, 11, 12 show the distributions of the mental workload scores with a red line indicating the normal distribution and the black line fitting the data.

4.1. 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-dimensional. Table 6 shows the Cronbach's Alpha coefficients of the other two selected multidimensional mental workload assessment instruments (NASA-TLX and the WP) obtained by considering all the answers of students across all the topics and design conditions. A reliability coefficient greater than .70 is deemed acceptable for considering a scale being consistent measure of a construct. Therefore,

Table 6

Reliability of the multidimensional mental workload scales, namely the Nasa Task Load Index (without and with the pair-wise procedure) and the Workload Profile with sample size, related number of items in the scales and associated Cronbach's Alpha. Note: NASA-TLX

Instrument	Sample size	Number of items	Cronbach's Alpha
The NASA-TLX (without pair-wise)	175	6	0.748
The NASA-TLX (with pair-wise)	175	6+15=21	0.612
The Workload Profile	173	8	0.876

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 research study. To confirm the obtained high reliability, Cronbach's Alpha has been computed also for each topic as well as each design condition. Table 7 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.

4.2. Validity

To assess the validity of the three selected mental workload assessment instruments, two sub-forms of validity were

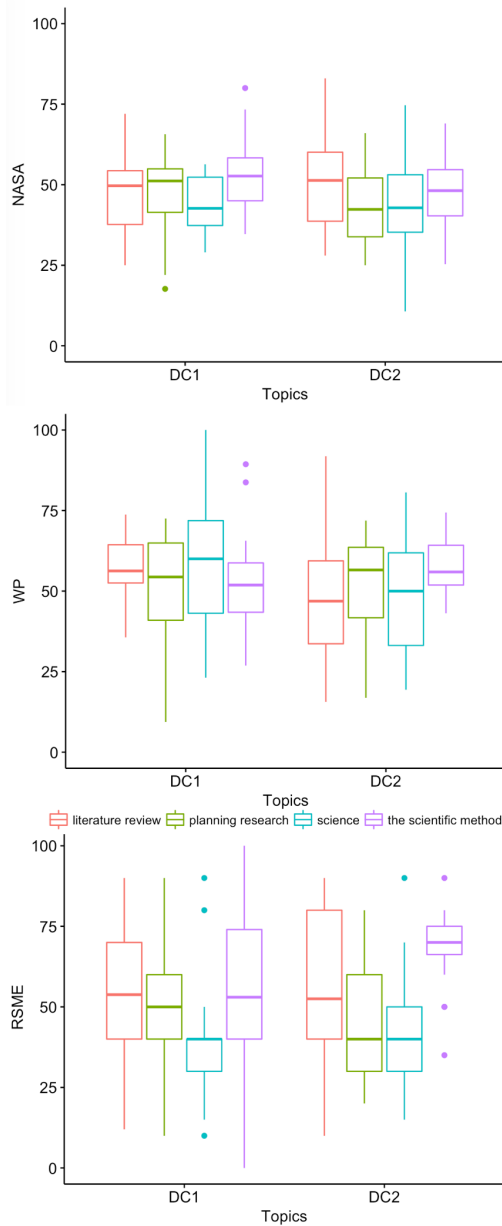


Figure 5: Boxplots of the distribution of the mental workload scores grouped by assessment instrument, design condition and topic

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 mental workload has been asked to students right before the completion of each topic and before the completion of the mental workload questionnaires. This question can be found in 9 and its answers have been correlated to the mental workload scores of the three selected measurement techniques. To assess convergent validity, the mental workload scores produced by the NASA-TLX and the Workload Profile in-

Table 7

Reliability of the multidimensional mental workload scales, namely the Nasa Task Load Index and the Workload Profile, grouped by topic, class and design condition (overall 17 classes

Topic	Mental Workload technique			
	NASA-TLX		WP	
	Size	Cronbach's Alpha	Size	Cronbach's Alpha

Design condition 1

T1 - Science	7	0.72	11	0.94
T1 - Science	6	0.53	6	0.86
T2 - The scientific method	10	0.68	13	0.89
T2 - The scientific method	11	0.66	10	0.84
T3 - Planning research	11	0.59	9	0.93
T3 - Planning research	11	0.87	10	0.93
T4 - Literature Review	10	0.43	11	0.88
T4 - Literature Review	11	0.86	10	0.22

Design condition 2

T1 - Science	13	0.85	13	0.82
T1 - Science	10	0.85	11	0.74
T1 - Science	13	0.70	9	0.95
T2 - The scientific method	12	0.45	12	0.6
T2 - The scientific method	6	0.76	6	0.79
T3 - Planning research	11	0.76	11	0.83
T3 - Planning research	11	0.57	11	0.48
T4 - Literature Review	13	0.81	11	0.92
T4 - Literature Review	9	0.91	9	0.95

struments have been correlated against the mental workload scores associated to the Rating Scale Mental Effort self-reporting uni-dimensional measure. 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 two instruments at the same time. Both the Pearson correlation coefficient and the Spearman's Rank correlation coefficient have been employed for computing the two forms of validity. Table 8 shows the correlations for face validity while table 9 lists the correlations for convergent validity.

4.3. Sensitivity

The sensitivity of the mental workload instruments has been calculated performing an analysis of the variance of the mental workload scores for each instrument. Figure 5 shows the boxplots of the mental workload scores associated to the three different mental workload measures grouped by topic, design condition and assessment technique. From a visual inspection of these box-plots, no difference between the mental workload scores associated to each topic when compared across the two design con-

Table 8

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

Instrument	Sample size	Pearson r	Spearman ρ
NASA-TLX	175	0.47	0.48
WP	173	0.46	0.45
RSME	348	0.44	0.43

Table 9

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

Instrument	Sample size	Pearson r	Spearman ρ
NASA-TLX vs RSME	175	0.51	0.49
WP vs RSME	173	0.37	0.39

ditions emerged. To confirm this, a formal comparison has been conducted to verify whether the distributions of the mental workload scores for each topic are statistically significant different across the two design conditions. Independent two-sample T-Tests (t) [16] 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 [64, 85].

All the p-values associated to the T -tests were greater than 0.05, therefore it is possible to conclude that the means of the two groups under comparison are not statistically significantly different. Similarly, Also the p-values associated to the V -tests are greater than 0.05, thus it is possible to conclude that the means have remained essentially unchanged. These comparisons can be visually inspected in the overlapping density plots of tables 13-16. From these it is possible to see that clear discriminations between the two design conditions cannot drawn.

These findings suggest 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. However, t-Test in general makes very strong assumptions: i) it is sufficient to detect changes in location; ii) the typical observation within each distribution can be summarised only by the mean; iii) the underlying compared distributions differ only in central tendency, not in other aspects. As it often happens in educational settings, sample size of sample students are very small, exactly as in this research. Additionally, there is no reason only to assume that the two distributions for each design condition differ only in the location of the set of the observations [82]. In fact, effects can be spotted in the tails of these distributions. For these considerations, a powerful descriptive statistics has been used to assess in details the differences between de-

sign conditions: shift functions [83]. Originally proposed in [15], a more systematic way to examine how two independent distributions differ exist. This includes plotting the difference between the quantiles associated to the two distributions of two groups as a function of the quantiles of one group. This version was further improved by Wilcox [81] in terms of better probability coverage and more power. This technique includes the use of the Harrell-Davis quantile estimator [84] and it computes confidence intervals of the decile differences with a bootstrap estimation of the standard error of the deciles. It is robust as it controls for multiple comparisons so that the type I error rate remains around 0.05 across the 9 confidence intervals. This leads to confidence intervals to be a bit larger than what they would be if only one decile was compared, thus the long-run probability of a type I error across all 9 comparisons remains near 0.05. This approach can be applied not only to deciles, but also to quartiles [63]. Figures 27, 28 and 29 present the Harrel-Davis quantile estimations of the mental workload scores respectively for the NASA-TLX, Workload Profile and the RSME instruments grouped by taught topic (as listed in figure 4).

Each plot depicts the differences between the two design conditions (DC1, DC2) for the quartiles. The three vertical lines in each plot indicate the confidence interval for each quartile (3 overall), with a dot in the middle, representing the mean. If this dot is above the zero horizontal line (null differences between the two design conditions) it means DC1 is more right-shifted than DC2. In other words, the mental workload scores in that quartile are higher for DC1. Contrarily, if the lines are below the zero line, it means DC2 is more right shifted than DC1, meaning that the mental workload scores for that quartile are higher for DC2. Yellow lines (and dots) underline the right shift of the values for DC1 for that quartile, whereas violet lines (and dots) for DC2. If a vertical confidence interval line does not cross the zero horizontal line, then the differences in mental workload scores for that quartile are considered significant in a frequentist term. These plots allow a deeper qualitative investigation of the differences between the two design conditions and a more detailed investigation of the sensitivity of the mental workload assessment instruments.

Figure 6 summarises graphically the differences and their magnitude between the two design conditions and the direction of these, grouped by mental workload assessment instrument and taught topic. Differences above the zero-line are higher for DC1, while below are in favour of DC2. Results are mixed and are further summarised in Figure 7 that depicts the number of null, weak and strong differences spotted by each mental workload assessment instrument. Differences are considered weak when the vertical line for a quartile crosses the zero line, but the mean is either above or below. Strong differences occur when a vertical line does not cross the zero-line, and so the mean is not on that, whereas null differences occur when the dot in the vertical line is on the zero horizontal line. The Workload

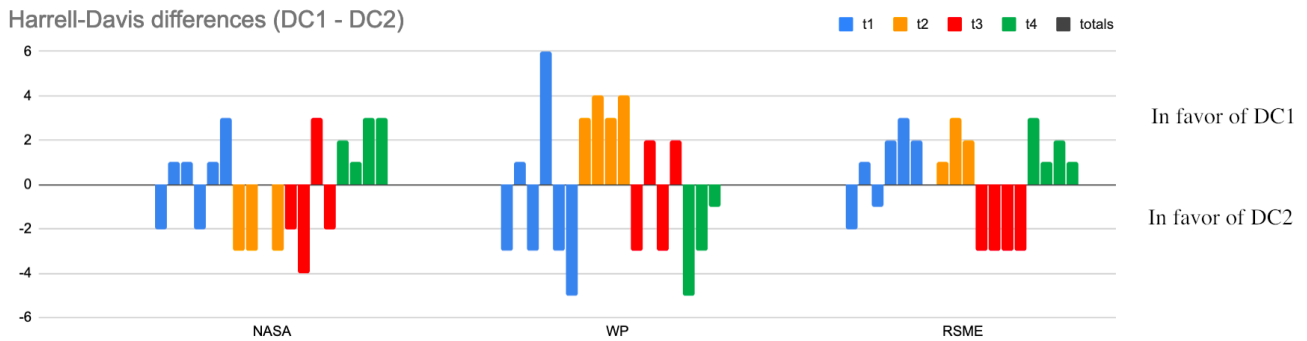


Figure 6: Harrell-Davis quartile differences between design conditions grouped by mental workload instrument

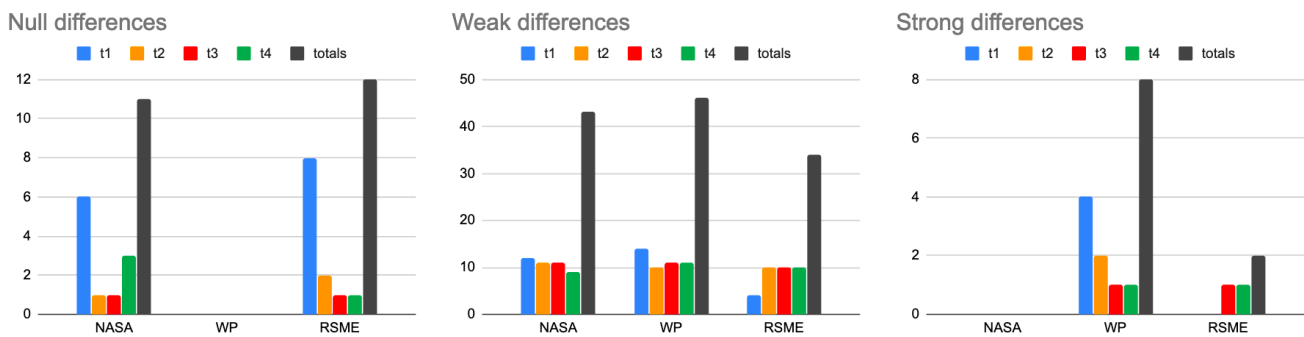


Figure 7: Details of Harrell-Davis quartile differences grouped by strength (null, weak, strong) and mental workload instrument

Profile instrument was the most sensitive to variations of the mental workload assessed for students, with overall 8 strong differences spotted against only 2 for the RSME and none for the NASA Task Load Index. All the three instruments, were able to spot many of the weak differences. The unidimensional RSME seem to be the most insensitive to mental workload variations followed by the NASA Task Load index. Therefore, the proportional of attentional resources used for attending a long university learning session, self-reported by students with the Workload profile instrument, seem to be a more suitable way to assess their mental workload than only reporting experienced effort via the RSME, or by using the NASA-TLX. Future scholars are encouraged to further explore these mental workload assessment instruments with other learning sessions, and test their suitability with long tasks.

4.4. Discussion

One uni-dimensional and two multidimensional self-reporting mental workload (MWL) assessment techniques, largely used within Ergonomics, have been employed in a novel primary research experiment within Education. The former is the Rating Scale Mental Effort [94] while the latter are the Nasa Task Load Index [22] and the Workload Profile [73]. These instruments, widely used within psychology, have been applied in a typical university classroom in the context of a module taught in the School of Computer Science, at the (XXX university, blind review). The exper-

iment involved the quantification of the mental workload experienced by students during third-level classes. Learners were exposed to two different design conditions across four topics of the post-graduate module ‘Research Design and proposal writing’. The former condition included the delivery of the topics by employing a traditional lecturer-to-students direct 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 the set of principles offered by the Cognitive Theory of Multimedia Learning [40].

An analysis of the reliability of the two multidimensional MWL assessment techniques has been conducted through a measure of their internal consistency. In detail, the Cronbach’s Alpha has been employed to assess the relation of the items associated to each technique. An obtained overall alpha value of 0.748 for the NASA task Load Index, considering all the students and topics, suggested that all its items share high covariance and probably measure the underlying construct (mental workload). This value dropped to 0.612 when the pairwise comparison, part of the instrument, was added to the reliability test. The situation is similar for the Workload Profile with an even higher alpha of 0.876 considering all students across all topics. 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 0.60 and 0.8 is required for reliability. Having reasonably reliable multidimensional measures of mental workload, an analysis of their validity has been subsequently performed. In particular, two forms of validity were employed: face and convergent validity. The former validity indicates the extent to which the three employed mental workload 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 of mental workload. The obtained Pearson and Spearman coefficients suggest how the three MWL measures are moderately correlated to a subjective value of overall mental workload self-reported by students, 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 achieved, the sensitivity of the three measures of mental workload 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 initially assessed through an analysis of the variance of the mental workload scores, associated to the four topics across the two design conditions, and a formal comparison of their distributions using the T-test or the Wilcoxon test, depending on their normality of the scores and other assumptions. Evidence suggests how the two design conditions imposed on average similar mental workload on students as computed using the three mental workload assessment techniques in terms of central tendency. Intuitively, given the strong reliability and moderate validity achieved by the selected mental workload measures, it seems to be reasonable to infer that the design principles from the Cognitive Theory of Multimedia Learning when applied for the design of the second instructional condition, were as not effective as expected because of the insensitivity to discriminate the two design conditions in this primary research.

However, a number of considerations are needed. Firstly, the low sensitivity of the mental workload measures (NASA, WP, RSME) adopted in this research to discriminate the two design conditions might be attributed to a particular line of thought emerged in the literature. This line of thought argues that despite the apparent advantage of presenting the learning material using auditorial and pictorial content embedded in a multimedia video, multimedia materials still require high levels of cognitive processing to synthesise the visual and auditory streams and extract the semantics of the main new information [26]. In fact, in this research, even if the second design condition resulted in the development of videos that led to significantly shorter

classes (as per table 4), even if the signalling principle was used to emphasise essential material with cues and the redundancy principle applied by removing most of the text, offloading one channel (eyes) and presented using a temporal alignment between words (verbally transmitted) and pictures, still students likely experienced high levels of cognitive processing for new information as no pre-training was offered to them. Secondly, the low sensitivity of the mental workload measures could be explained by the fact that both the design conditions were adhering already to the coherence principle, with extraneous material kept to minimum, and to the spatial contiguity principle, by which words and pictures were showed at the same time (slide or screen), using the same conversational style (personalisation principle), narrated by the lecturer and not by an artificial machine synthesiser (voice principle). This overlapping might explain the non-significant difference of the obtained mental workload scores across the two design conditions, which were on average in the middle range of the mental workload distributions, that means no situations of underload or overload. Similarly, from the data in table 2, it is possible to note that the CTML principles associated to the germane load type of CLT (multimedia, personalisation, voice principles, as described in table 1) have not been altered across the two design conditions, thus, with high probability, germane load was not differently promoted. This is gauged by the perceived effort rated by students, using the Rating Scale Mental Effort (RSME) workload instrument, that resulted in scores not significantly different across the two design conditions.

As suggested by [36], there is uncertainty and dynamism behind his Cognitive Theory of Multimedia Learning, suggesting how its principles should not be taken rigidly, but rather as a starting point for discussion and experimentation, as showed in this research. Multimedia videos have the high potential to improve learning in various ways, however they can also be very demanding for the cognitive system of a learner. The findings from the present study are aligned to previous research on CTML and the underlying assumption driven from it whereby learners and their mind can process only small portions of large amount of auditory and visual stimuli at one time [35, 42]. The experiment proposed in this study can be seen as a potential solution to the problem highlighted by [32], whereby, nowadays, the majority of faculty members are involved in instructional design activities that mostly lack scientific underpinning and proper documentation, therefore more evidence-based instructional designs are needed. Eventually, this experiment is in line with the critics carried forward by [9] whereby multimedia instruction is an example of a new area of instructional research and practice that has led to a significant amount of excitement among educational scholars. Like any new area of research, CTML is based upon a set of assumptions about the way students learn and will solve problems, therefore implicitly creating a set of expectations about multimedia benefits. However,

these assumptions and expectations are repeatedly taken for granted, valid and appropriate, often forgetting about their empirical validation through evidence-based research. Clark et al. insist that if these implicit assumptions turn out to be incorrect, researchers may unintentionally adopt them for designing their multimedia instructional material that does not support learning [9]. Additionally, when a new research finding is in contradiction to a previous one, for instance as demonstrated in this research, there is a tendency among scholars to ignore it by simply justifying this as a poorly designed multimedia video rather than performing a careful analysis.

For the above reasons, a deeper sensitivity analysis has been performed on the mental workload scores obtained with experimental research by applying the Harrell-Davis estimator [21]. This is a weighted linear combination of order statistics in which the order statistics used in traditional non-parametric quantile estimators are given the greatest weight. It is suitable for small sample sizes, as those associated to the university classes of this research (between 7 and 15 students per group) and it represents the limit of a bootstrap average as the number of bootstrap resamples becomes infinitely large. This estimator represents a more systematic way to characterise how two independent distributions differ not only in central tendency, but also in their tails. It is a descriptive statistics that was applied using quartiles and that actually revealed many differences between the two design conditions. Results showed that the Workload Profile mental workload assessment instrument was the most sensitive to mental workload changes, followed by the NASA Task Load Index and the Rating Scale Mental Effort. Also, the fact that the RMSE has low sensitivity can be explained by the results obtained in [75] whereby authors have demonstrated that timing and frequency of effort ratings influence results and that repeated measures of mental effort, especially in long sessions, are preferable.

This research, through the adoption of quantitative measures of mental workload, can be seen as a way to provide scholars with a new set of mental workload measures that can be used to enhance the empirical validation of their instructional designs and facilitate comparisons across research studies. Additionally, in case of small sample sizes, as it often happens within education, and when effects are not necessarily homogenous among participants, the application of the Harrell-Davis estimator is very useful to determine how, and to what extent, two distributions differ.

5. Conclusions

This study attempted to investigate the impact of three mental workload assessment techniques, on the evaluation of different instructional design conditions. A primary research study has been performed in a typical university classroom and a case study involved the development and evaluation of two design conditions. The former

condition included the delivery of four topics by employing the traditional direct instructions method lecturer-driven whereby the delivery of instructional material was done by 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 the set of principles proposed within the Cognitive Theory of Multimedia Learning [40]. Empirical evidence strongly suggested how the three MWL measures are reliable when applied in a typical university classroom. Results demonstrated their moderate validity, in line with the validity achieved in other experiments within ergonomics and psychology. On the contrary, their sensitivity was very low in discriminating the two instructional design conditions. However, given the high reliability and modest validity of the three MWL measures, the achieved sensitivity might reasonably underline the minimal impact of the principles of Cognitive Theory of Multimedia Learning for developing the second design condition, in line with other research studies [26, 35, 42].

The contributions of this research are to offer a replicable methodology for the evaluation and application of existing mental workload measures in education which in turn supports empirical and evidence-based instructional design. Contrarily to the limited falsifiability of Cognitive Load Theory, mainly given by the robustness of measures for its cognitive load types, as reported in the literature, this study conforms to the rules of science because its methodology is replicable and its adoption might lead to new findings, falsifying existing ones. Every single test of existing measures of mental workload in Education is aimed at increasing our understanding of mental workload itself as a construct, and in turn facilitating the exploration of the effectiveness of various instructional designs all aimed at enhancing learning.

Future work might include the application of more advanced principles of CTML [39] for the development of various design conditions. These might include, for instance, the application of the navigation principle by which humans learn better in those instructional environments where relevant navigational aids are provided or the application of the collaborative principle, by which people learn better when involved in collaborative learning activities. Additionally, further multidimensional and unidimensional measures of mental workload might be employed and tested against reliability, validity and sensitivity. The multimedia artefacts developed in this research might be also extended as suggested in [3] by adopting guiding questions. These are questions available upfront and during the entire multimedia video and, they represent a means to share learning objectives with students, aimed at promoting their germane load devoted to the underlying learning task and reducing their extraneous load by redirecting their attention to the core instructional elements. Eventually, other instructional designs conditions, should be designed and tested against mental

workload. In particular, following the recent suggestions for complex learning environments [31], more flexible approaches to learning based on differentiating specific goals of various learning activities can be tested. In details, mental workload assessment at various instructional stages and tasks can lead to more precise findings as these stages and tasks might be linked to different specific goals. Eventually, the three mental workload assessment instruments used in this research can be employed with existing measures of efficiency [29, 76, 49, 50] and engagement [25], extending current evaluations of instructional designs.

References

- [1] Paul Ayres. Using subjective measures to detect variations of intrinsic cognitive load within problems. *Learning and Instruction*, 16(5):389–400, 2006.
- [2] Alan Baddeley and Graham James Hitch. Working memory. In G.A. Bower, editor, *Recent Advances in Learning and Motivation*, volume 8, pages 47–90. Academic Press, 1974.
- [3] Cynthia J Brame. Effective educational videos: Principles and guidelines for maximizing student learning from video content. *CBE-Life Sciences Education*, 15(4):es6, 2016.
- [4] Karel A Brookhuis and Dick de Waard. Monitoring drivers' mental workload in driving simulators using physiological measures. *Accident Analysis & Prevention*, 42(3):898–903, 2010.
- [5] Roland Ed Brünken, Jan L Plass, and Roxana Ed Moreno. Current issues and open questions in cognitive load research. 2010.
- [6] Brad Cain. A review of the mental workload literature. Technical report, Defence Research and Development Canada Toronto, Human System Integration Section, 2007.
- [7] P. Chandler and J. Sweller. Cognitive load theory and the format of instruction. *Cognition and Instruction*, 8(4):293–332, 1991.
- [8] Gabriele Cierniak, Katharina Scheiter, and Peter Gerjets. 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.
- [9] Richard E. Clark and David F. Feldon. Five common but questionable principles of multimedia learning. *The Cambridge handbook of multimedia learning*, 6, 2005.
- [10] Ton De Jong. Cognitive load theory, educational research, and instructional design: some food for thought. *Instructional science*, 38(2):105–134, 2010.
- [11] Nicolas Dehue and Cécile van de Leemput. What does germane load mean? an empirical contribution to the cognitive load theory. *Frontiers in Psychology*, 5:1099, 2014.
- [12] Krista E DeLeeuw and Richard E Mayer. 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] Angela DiDomenico and Maury A Nussbaum. Interactive effects of physical and mental workload on subjective workload assessment. *International journal of industrial ergonomics*, 38(11-12):977–983, 2008.
- [14] Peter Dixon. From research to theory to practice: Commentary on chandler and sweller. *Cognition and Instruction*, 8(4):343–350, 1991.
- [15] Kjell Doksum. Empirical probability plots and statistical inference for nonlinear models in the two-sample case. *The annals of statistics*, pages 267–277, 1974.
- [16] Joan Fisher Box. Guinness, gosset, fisher, and small samples. *Statistical Science*, 2(1):45–52, 1987.
- [17] Peter Gerjets, Katharina Scheiter, and Gabriele Cierniak. 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.
- [18] Susan R Goldman. On the derivation of instructional applications from cognitive theories: Commentary on chandler and sweller. *Cognition and Instruction*, 8(4):333–342, 1991.
- [19] Jacek Gwizdzka. Assessing cognitive load on web search tasks. *The ergonomic open journal*, 2(1):114–123, 2009.
- [20] Jacek Gwizdzka. Distribution of cognitive load in web search. *Journal of the american society & information science & technology*, 61(11):2167–2187, November 2010.
- [21] Frank E Harrell and CE Davis. A new distribution-free quantile estimator. *Biometrika*, 69(3):635–640, 1982.
- [22] Sandra G. Hart. 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, 2006. Sage Journals.
- [23] Sandra G. Hart. Nasa-task load index (nasa-tlx); 20 years later. *Proceedings of the human factors and ergonomics society annual meeting*, 50(9):904–908, 2006.
- [24] Sandra G Hart and Lowell E. Staveland. Development of nasa-tlx (task load index): results of empirical and theoretical research. In Peter A. Hancock and Najmedin Meshkati, editors, *Human Mental Workload*, volume 52 of *Advances in Psychology*, pages 139–183. North-Holland, 1988.
- [25] Curtis R Henrie, Lisa R Halverson, and Charles R Graham. Measuring student engagement in technology-mediated learning: A review. *Computers & Education*, 90:36–53, 2015.
- [26] B. D. Homer, J. L. Plass, and L. Blake. The effects of video on cognitive load and social presence in multimedia-learning. *Computers in Human Behavior*, 24(3):786–797, 2008.
- [27] Beverly Messick Huey and Christopher D. Wickens. *Workload transition: implication for individual and team performance*. National Academy Press, Washington, DC., 1993.
- [28] Dayu Jiang and Slava Kalyuga. Confirmatory factor analysis of cognitive load ratings supports a two-factor model. *Tutorials in Quantitative Methods for Psychology*, 16:216–225, 2020.
- [29] Slava Kalyuga. Enhancing instructional efficiency of interactive e-learning environments: A cognitive load perspective. *Educational psychology review*, 19(3):387–399, 2007.
- [30] Slava Kalyuga. Cognitive load theory: How many types of load does it really need? *Educational Psychology Review*, 23(1):1–19, 2011.
- [31] Slava Kalyuga and Anne-Marie Singh. Rethinking the boundaries of cognitive load theory in complex learning. *Educational Psychology Review*, 28(4):831–852, 2016.
- [32] A.J. Levinson. Where is evidence-based instructional design in medical education curriculum development? *Medical Education*, 44:536–537, 2010.
- [33] Luca Longo. *Formalising Human Mental Workload as a Defeasible Computational Concept*. Doctor in philosophy, School of Computer Science and Statistics - Trinity College Dublin, 2014.
- [34] Luca Longo and Pierpaolo Dondio. On the relationship between perception of usability and subjective mental workload of web interfaces. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2015 IEEE/WIC/ACM International Conference on*, volume 1, pages 345–352. IEEE, 2015.
- [35] P.D. Mautone and R.E Mayer. Signaling as a cognitive guide in multimedia learning. *Journal of Educational Psychology*, 93(2):377, 2001.
- [36] R.E. Mayer. Using multimedia for e-learning. *Journal of Computer Assisted Learning*, 33(5):403–423, 2017. JCAL-16-266.R1.
- [37] Richard E Mayer. Multimedia learning: Are we asking the right questions? *Educational psychologist*, 32(1):1–19, 1997.
- [38] Richard E. Mayer. Multimedia learning. *Psychology of Learning and Motivation*, 41:85–139, 2002.
- [39] Richard E Mayer. *The Cambridge handbook of multimedia learning*. Cambridge university press, 2005.
- [40] Richard E Mayer. *Multimedia learning*. Cambridge University Press, 2009.
- [41] Richard E. Mayer. Information processing. pages 85–99. American Psychological Association, 2012.
- [42] Richard E Mayer and Roxana Moreno. Nine ways to reduce cognitive load in multimedia learning. *Educational psychologist*, 38(1):43–52,

- 2003.
- [43] G. A. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63(2):81–97, 1956.
- [44] Brian W Moroney, Joel S Warm, and William N Dember. Effects of demand transitions on vigilance performance and perceived workload. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 39(21):1375–1379, 1995.
- [45] S. Mousavi, R. Low, and J. Sweller. Reducing cognitive load by mixing auditory and visual presentation modes. *Journal of Educational Psychology*, 87(2):319–334, 1995.
- [46] Karim Moustafa and Luca Longo. Analysing the impact of machine learning to model subjective mental workload: A case study in third-level education. In Luca Longo and M. Chiara Leva, editors, *Human Mental Workload: Models and Applications*, pages 92–111, Cham, 2019. Springer International Publishing.
- [47] Karim Moustafa, Saturnino Luz, and Luca Longo. Assessment of mental workload: a comparison of machine learning methods and subjective assessment techniques. In *International Symposium on Human Mental Workload: Models and Applications*, pages 30–50. Springer, 2017.
- [48] Jan M Noyes and Daniel PJ Bruneau. A self-analysis of the nasa-tlx workload measure. *Ergonomics*, 50(4):514–519, 2007.
- [49] Giuliano Orru, Federico Gobbo, Declan O’Sullivan, and Luca Longo. 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.*, pages 292–302, 2018.
- [50] Giuliano Orru and Luca Longo. Direct instruction and its extension with a community of inquiry: A comparison of mental workload, performance and efficiency. In *Proceedings of the 11th International Conference on Computer Supported Education, CSEDU 2019, Heraklion, Crete, Greece, May 2-4, 2019, Volume 1.*, pages 436–444, 2019.
- [51] Giuliano Orru and Luca Longo. The evolution of cognitive load theory and the measurement of its intrinsic, extraneous and germane loads: A review. In Luca Longo and M. Chiara Leva, editors, *Human Mental Workload: Models and Applications*, pages 23–48, Cham, 2019. Springer International Publishing.
- [52] Fred Paas, Juhani E Tuovinen, Huib Tabbers, and Pascal WM Van Gerven. Cognitive load measurement as a means to advance cognitive load theory. *Educational psychologist*, 38(1):63–71, 2003.
- [53] Fred Paas and Jeroen J. G. Van Merriënboer. 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.
- [54] A. Paivio. *Mental Representations: A Dual Coding Approach*. Oxford Psychology Series. Oxford University Press, 1990.
- [55] Karl Popper. *The logic of scientific discovery*. Routledge, 2005.
- [56] Karl Popper. *Conjectures and refutations: The growth of scientific knowledge*. routledge, 2014.
- [57] Sébastien Puma, Nadine Matton, Pierre-Vincent Paubel, and André Tricot. Cognitive load theory and time considerations: Using the time-based resource sharing model. *Educational Psychology Review*, 30(3):1199–1214, 2018.
- [58] Gary B. Reid and Thomas E. Nygren. The subjective workload assessment technique: A scaling procedure for measuring mental workload. In Peter A. Hancock and Najmedin Meshkati, editors, *Human Mental Workload*, volume 52 of *Advances in Psychology*, chapter 8, pages 185–218. North-Holland, 1988.
- [59] Lucas Rizzo, Pierpaolo Dondio, Sarah Jane Delany, and Luca Longo. Modeling mental workload via rule-based expert system: a comparison with nasa-tlx and workload profile. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 215–229. Springer, 2016.
- [60] Lucas Rizzo and Luca Longo. Inferential models of mental workload with defeasible argumentation and non-monotonic fuzzy reasoning: a comparative study. In *Proceedings of the 2nd Workshop on Advances In Argumentation In Artificial Intelligence, co-located with XVII International Conference of the Italian Association for Artificial Intelligence, AI3@AI*IA 2018, 20-23 November 2018, Trento, Italy*, pages 11–26. CEURS, 2018.
- [61] Lucas Middeldorf Rizzo and Luca Longo. 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 AI3@AI*IA*. CEURS, 2017.
- [62] Alan H. Roscoe and George. A. Ellis. A subjective rating scale for assessing pilot workload in flight: A decade of practical use. Technical report TR 90019, Royal Aerospace Establishment, Farnborough (UK), March 1990.
- [63] Guillaume A Rousselet, Cyril R Pernet, and Rand R Wilcox. Beyond differences in means: robust graphical methods to compare two groups in neuroscience. *European Journal of Neuroscience*, 46(2):1738–1748, 2017.
- [64] Patrick Royston. Remark AS R94: A remark on algorithm AS 181: The W-test for normality. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 44(4):547–551, 1995.
- [65] Susana Rubio, Eva Diaz, Jesus Martin, and Jose M. Puente. Evaluation of subjective mental workload: A comparison of swat, nasa-tlx, and workload profile methods. *Applied Psychology*, 53(1):61–86, 2004.
- [66] Wolfgang Schnotz and Christian Kürschner. A reconsideration of cognitive load theory. *Educational Psychology Review*, 19(4):469–508, 2007.
- [67] Tina Seufert, Inge Jänen, and Roland Brünken. The impact of intrinsic cognitive load on the effectiveness of graphical help for coherence formation. *Computers in Human Behavior*, 23(3):1055–1071, 2007.
- [68] J. Sweller, J. Van Merriënboer, and F. Paas. Cognitive architecture and instructional design. *Educational Psychology Review*, 10(3):251–296, 1998.
- [69] John Sweller. Cognitive load theory, learning difficulty, and instructional design. *Learning and instruction*, 4(4):295–312, 1994.
- [70] John Sweller. Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational psychology review*, 22(2):123–138, 2010.
- [71] Andrew J Tattersall and Penelope S Foord. An experimental evaluation of instantaneous self-assessment as a measure of workload. *Ergonomics*, 39(5):740–748, 1996.
- [72] Pamela S. Tsang. Mental workload. In Waldemar Karwowski, editor, *Encyclopedia of Ergonomics and Human Factors*, volume 1, chapter 166. Taylor & Francis, 2006.
- [73] Pamela S. Tsang and Velma L. Velazquez. Diagnosticity and multidimensional subjective workload ratings. *Ergonomics*, 39(3):358–381, 1996.
- [74] Pamela S. Tsang and Michael A. Vidulich. Mental workload and situation awareness. In Gavriel Salvendy, editor, *Handbook of Human Factors and Ergonomics*, pages 243–268. Wiley & Sons, 2006.
- [75] Tamara Van Gog, Femke Kirschner, Liesbeth Kester, and Fred Paas. Timing and frequency of mental effort measurement: Evidence in favour of repeated measures. *Applied cognitive psychology*, 26(6):833–839, 2012.
- [76] Tamara Van Gog and Fred Paas. Instructional efficiency: Revisiting the original construct in educational research. *Educational Psychologist*, 43(1):16–26, 2008.
- [77] Michael A. Vidulich and Schueren James. Ward Frederic G. Using the subjective workload dominance (sword) technique for projective workload assessment. *Human Factors Society*, 33(6):677–691, December 1991.
- [78] Christopher D. Wickens. Multiple resources and mental workload. *Human Factors*, 50(2):449–454, 2008.
- [79] Christopher D. Wickens and Justin G. Hollands. *Engineering Psychology and Human Performance*. Prentice Hall, 3rd edition, September 1999.

- [80] Eric N. Wiebe, Edward Roberts, and Tara S. Behrend. An examination of two mental workload measurement approaches to understanding multimedia learning. *Comput. Hum. Behav.*, 26(3):474–481, 2010.
- [81] Rand R Wilcox. Comparing two independent groups via multiple quantiles. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 44(1):91–99, 1995.
- [82] Rand R Wilcox. Comparing two independent groups via a quantile generalization of the wilcoxon-mann-whitney test. *Journal of Modern Applied Statistical Methods*, 11(2):2, 2012.
- [83] Rand R Wilcox and David M Erceg-Hurn. Comparing two dependent groups via quantiles. *Journal of Applied Statistics*, 39(12):2655–2664, 2012.
- [84] Rand R Wilcox, David M Erceg-Hurn, Florence Clark, and Michael Carlson. Comparing two independent groups via the lower and upper quantiles. *Journal of Statistical Computation and Simulation*, 84(7):1543–1551, 2014.
- [85] Frank Wilcoxon. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6):80–83, December 1945.
- [86] Glenn F. Wilson and Thomas F. Eggemeier. Mental workload measurement. In Waldemar Karwowski, editor, *Int. Encyclopedia of Ergonomics and Human Factors (2nd ed.)*, volume 1, chapter 167. Taylor and Francis, 2006.
- [87] Glenn F Wilson and Christopher A Russell. Real-time assessment of mental workload using psychophysiological measures and artificial neural networks. *Human factors*, 45(4):635–644, 2003.
- [88] Bin Xie and Gavriel Salvendy. Review and reappraisal of modelling and predicting mental workload in single and multi-task environments. *Work and Stress*, 14(1):74–99, 2000.
- [89] Mark S. Young, Karel A. Brookhuis, Christopher D. Wickens, and Peter A. Hancock. State of science: mental workload in ergonomics. *Ergonomics*, 58(1):1–17, 2015.
- [90] Mark S. Young and Neville A Stanton. Mental workload. In Neville Anthony Stanton, Alan Hedge, Karel Brookhuis, Eduardo Salas, and Hal W. Hendrick, editors, *Handbook of Human Factors and Ergonomics Methods*, chapter 39, pages 1–9. CRC Press, 2004.
- [91] Mark S. Young and Neville A. Stanton. Mental workload: theory, measurement, and application. In Waldemar Karwowski, editor, *Encyclopedia of ergonomics and human factors*, volume 1, pages 818–821. Taylor & Francis, 2nd edition, 2006.
- [92] M.S. Young and N.A. Stanton. Mental workload: theory, measurement, and application. In *International encyclopedia of ergonomics and human factors*, volume 1, pages 507–509. London: Taylor and Francis, 2001.
- [93] Robert Z Zheng. *Cognitive load measurement and application: a theoretical framework for meaningful research and practice*. Routledge, 2017.
- [94] Ferdinand. R. H. Zijlstra. *Efficiency in work behaviour*. Doctoral thesis, Delft University, The Netherlands, 1993.

Table 10
The NASA Task Load Index mental workload assessment instrument

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 and 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 11
The Workload Profile mental workload assessment instrument

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)?

Please indicate, by marking the horizontal axis below, how much effort it took for you to attend the class.
Please indicate, by marking the horizontal axis below, how much effort it took for you to attend the class.

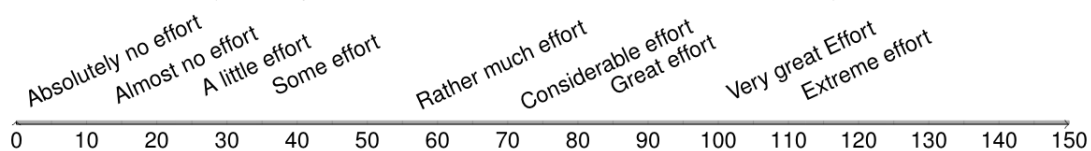


Figure 8: The Rating Scale Mental Effort assessment instrument

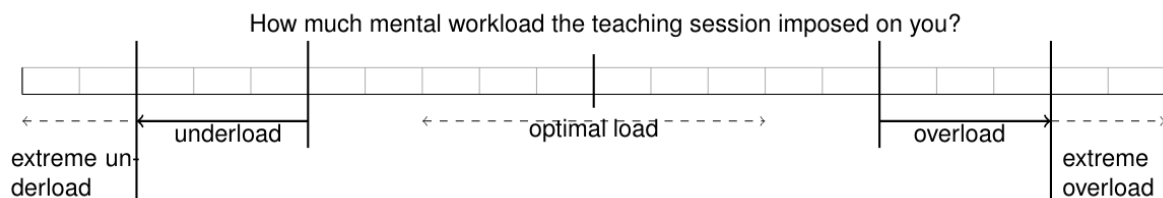


Figure 9: A self-reporting uni-dimensional measure of perception of mental workload

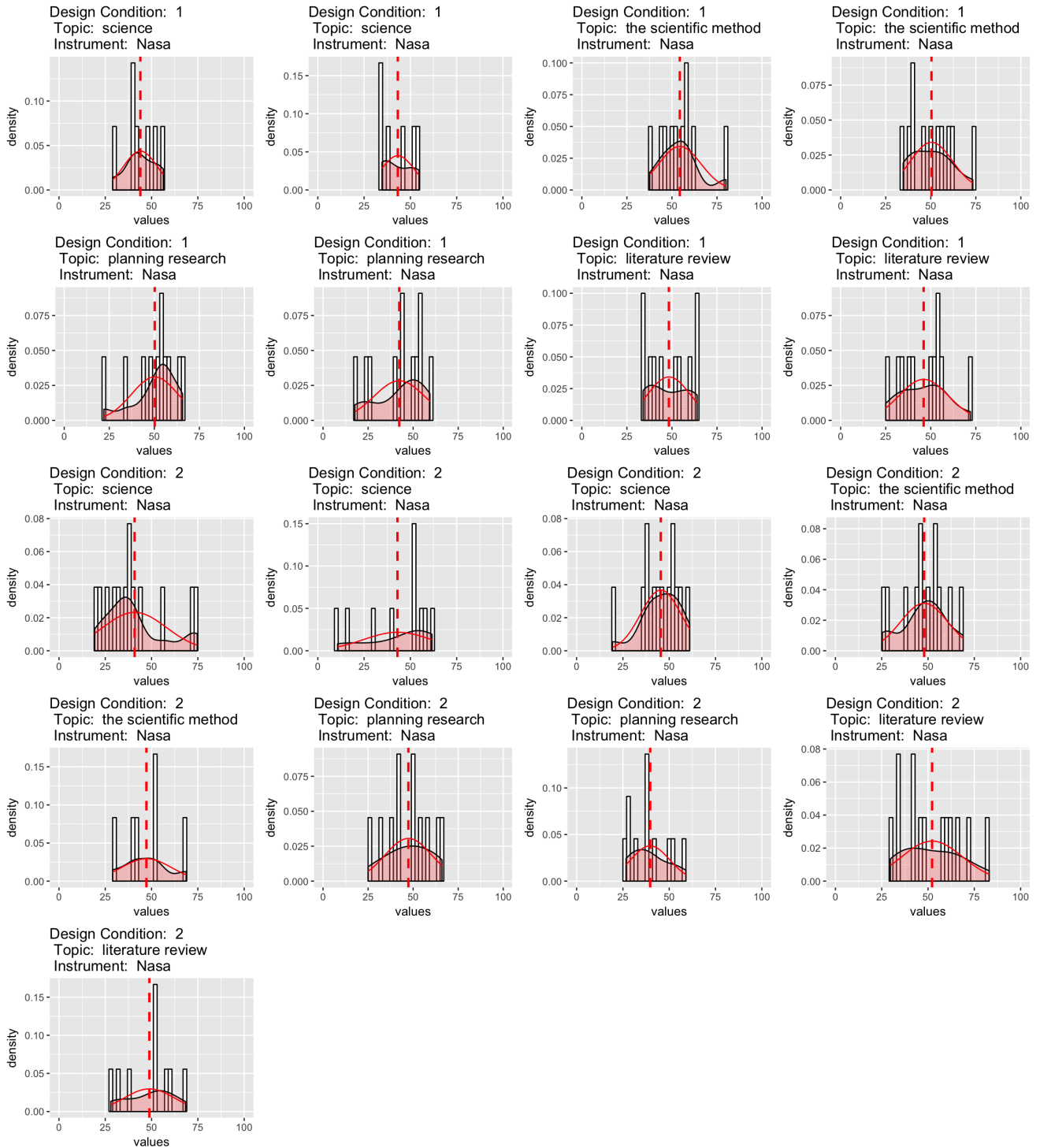


Figure 10: Distributions of the NASA-TLX scores grouped by design condition and topic

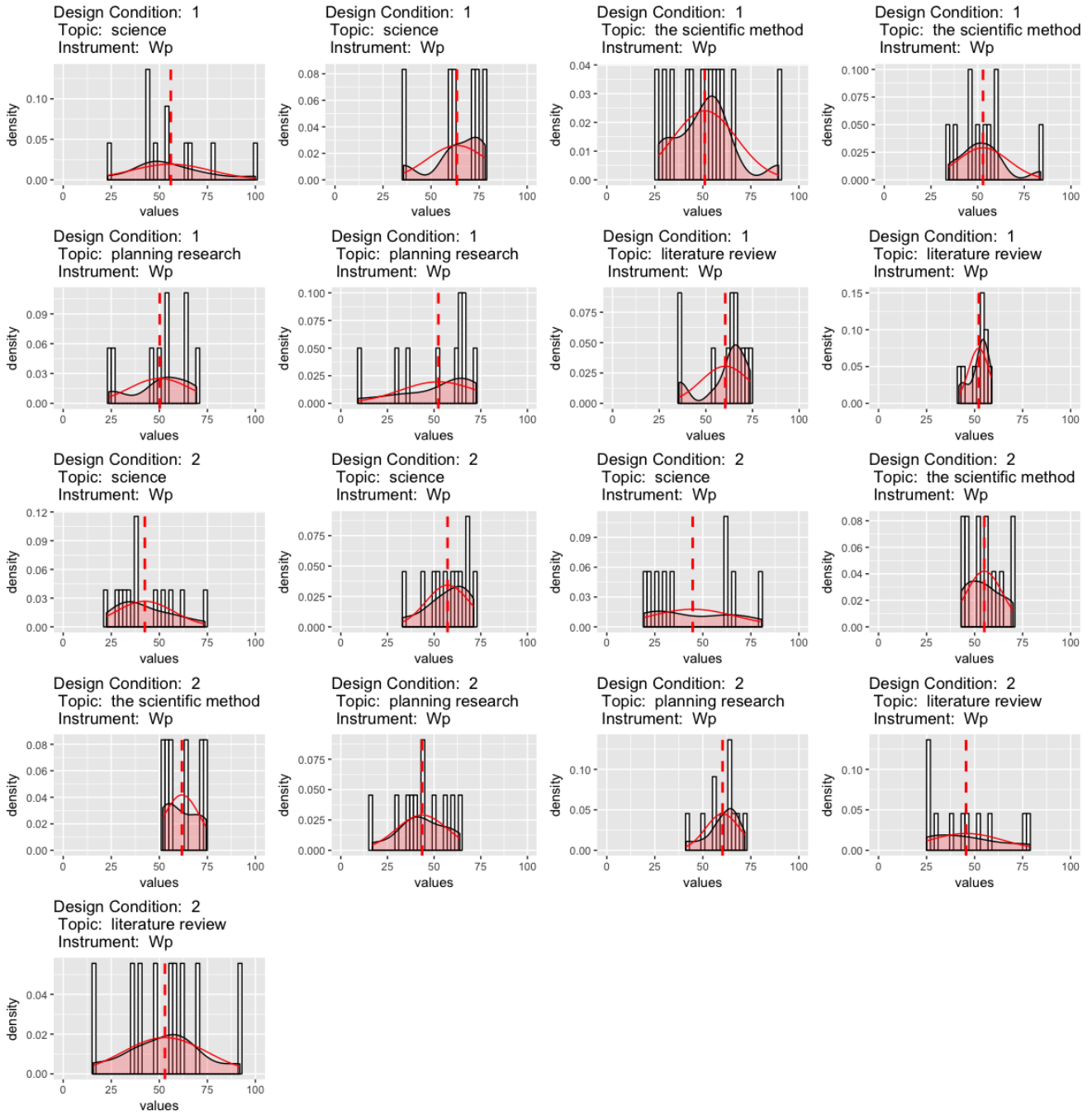


Figure 11: Distributions of the Workload Profile scores grouped by design condition and topic

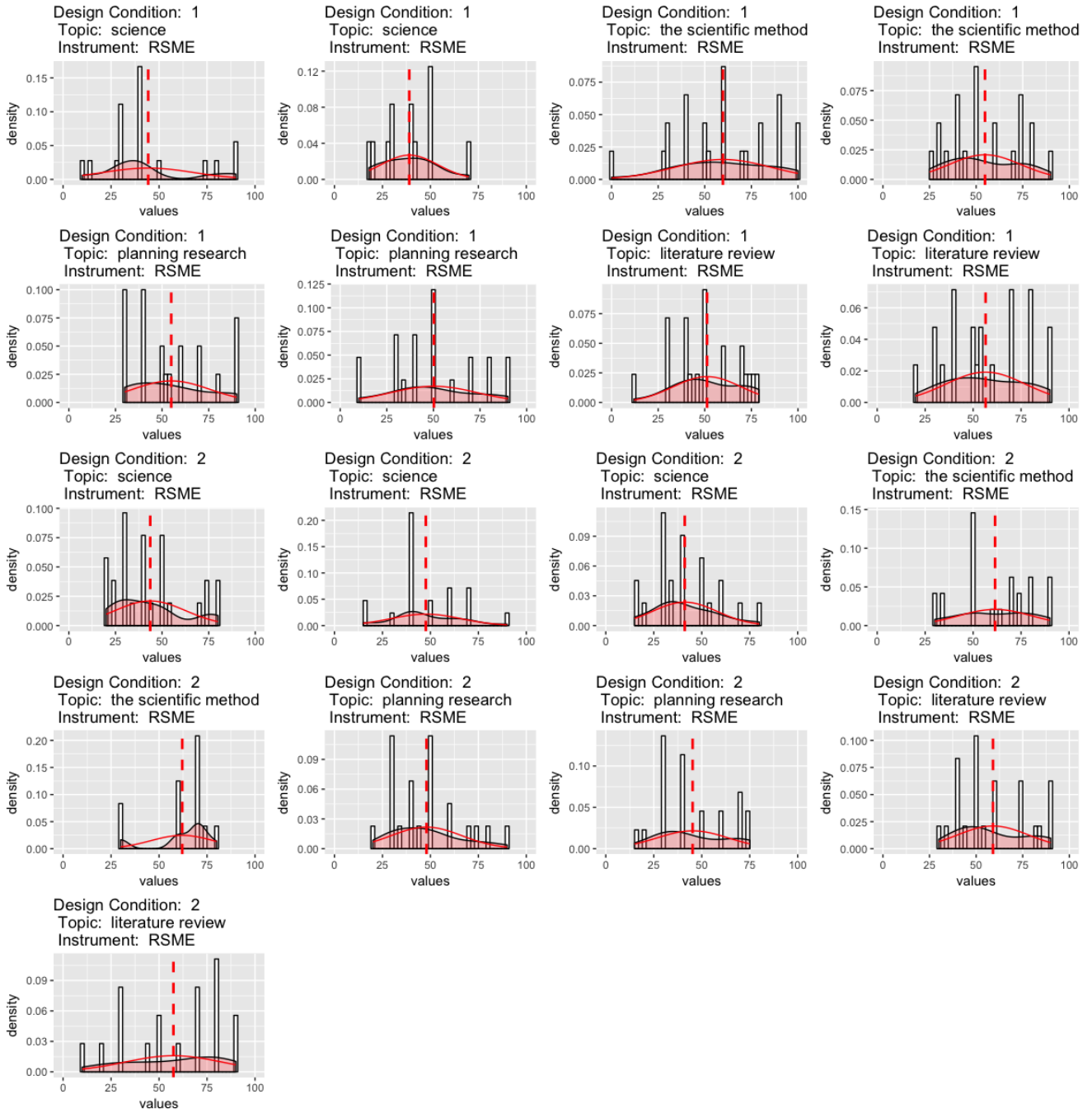


Figure 12: Distributions of the Rating Scale Mental Effort scores grouped by condition and topic

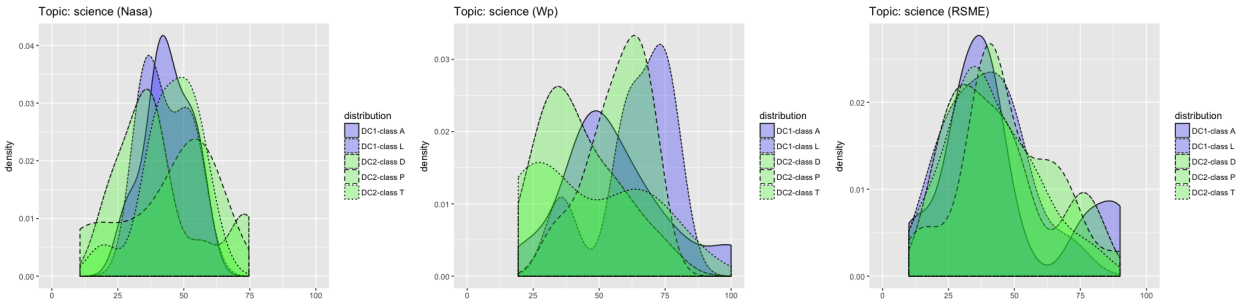


Figure 13: Comparison of the distributions of the mental workload scores for the topic 'science' grouped by instrument

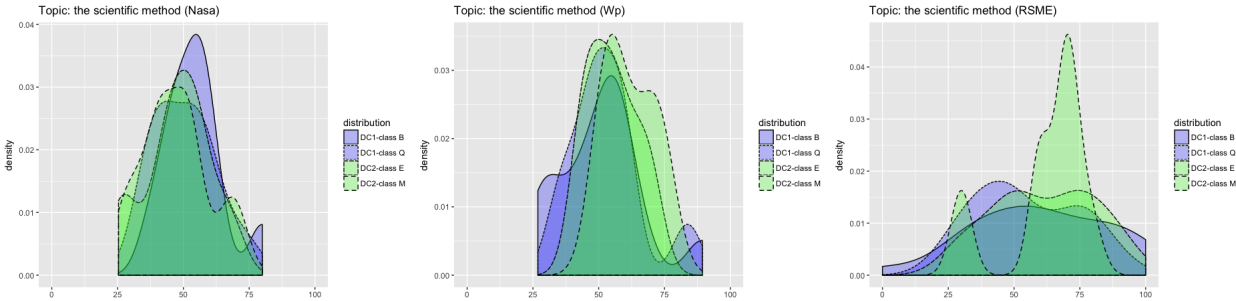


Figure 14: Comparison of the distributions of the mental workload scores for the topic 'the scientific method' grouped by instrument

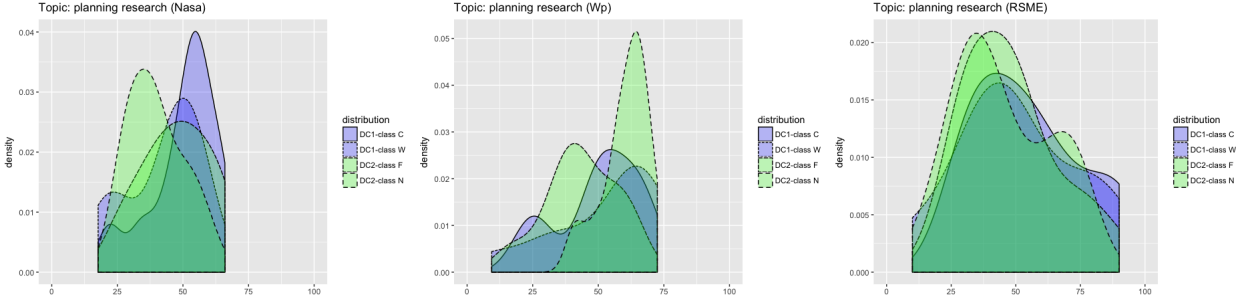


Figure 15: Comparison of the distributions of the mental workload scores for the topic 'planning research' grouped by instrument

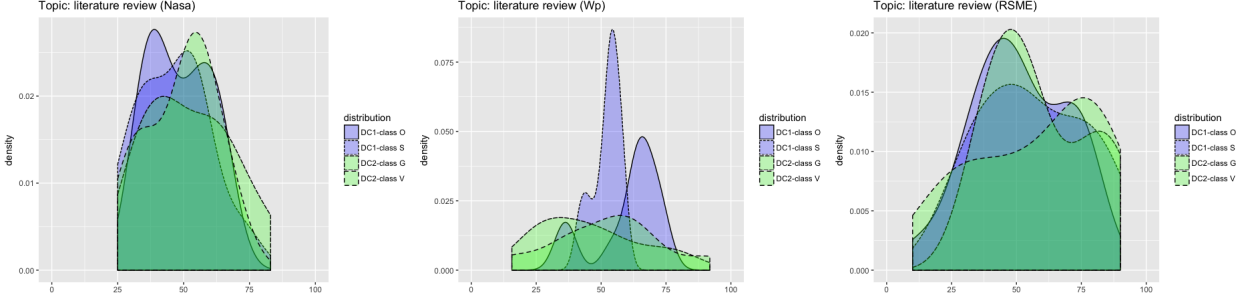


Figure 16: Comparison of the distributions of the mental workload scores for the topic 'literature review' grouped by instrument

Science

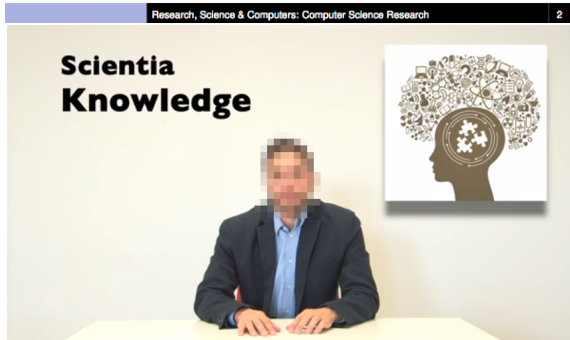
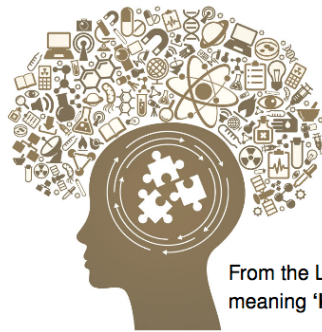


Figure 17: Application of the coherence principle of Cognitive Theory of Multimedia Learning in design condition 2 (bottom) as compared to design condition 1 (top)

Science - definitions ^{1 2}

- what is known, corpus of human **knowledge** (of something) acquired by study
- **knowledge**, learning, application
- is a systematic enterprise that builds and organises **knowledge** in the form of testable explanations and predictions about the universe
- refers to a body of **knowledge** that can be rationally explained and reliably applied

¹Online Etymology Dictionary - <http://www.etymonline.com/>
²Merriam-Webster: <http://www.merriam-webster.com/>

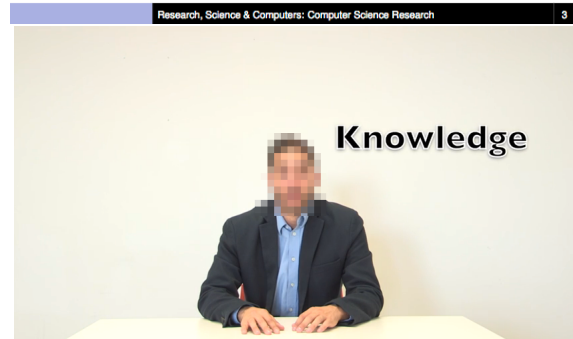


Figure 19: Application of the redundancy principle of Cognitive Theory of Multimedia Learning in design condition 2 (bottom) as compared to design condition 1 (top)

The scientific method - definitions 1 of 2

A **method** or procedure that has characterised natural science since the 17th century, consisting in **systematic observation, measurement, and experiment, and the formulation, testing, and modification of hypotheses** ¹

The **scientific method** is the process by which **science** is carried out.

¹<http://www.oxforddictionaries.com/definition/english/scientific-method>.

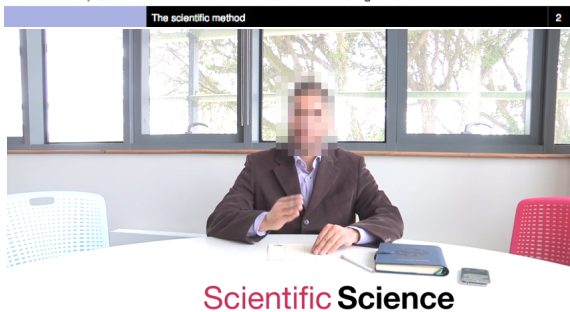


Figure 18: Application of the signaling principle of Cognitive Theory of Multimedia Learning in design condition 2 (bottom) as compared to design condition 1 (top)

Kuhn's paradigm

Kuhn's opinion of **science** was that, in fact, it is not a cumulative process, but in reality, a **cyclical process**.

A particular research perspective (paradigm) dominates for a period of time, until a new one is developed which supersedes it.

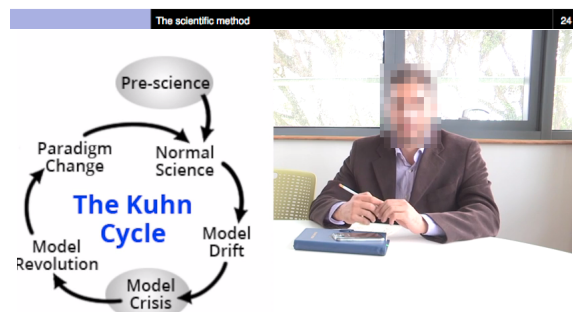
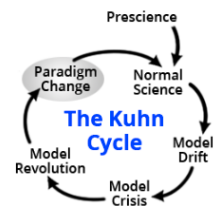


Figure 20: Application of the spatial contiguity principle of Cognitive Theory of Multimedia Learning in design condition 2 (bottom) as compared to design condition 1 (top)

Famous scientists - 9 of 9



- **name:** Enrico Fermi
- **born:** September 29th, 1901, Rome, Italy
- **died:** November 28th, 1954, Chicago, Illinois, United States
- **fields:** physics, statistical mechanics

Nobel prize, he is one of the men referred to as the **'father of the atomic bomb'** as the inventor of the **first nuclear reactor**. He was widely regarded as one of the very few physicists to excel both theoretically and experimentally.



Figure 21: Application of the temporal contiguity principle of Cognitive Theory of Multimedia Learning in design condition 2 (bottom) as compared to design condition 1 (top)

Literature review - final remark - 2 of 3

It's a bad sign to see every paragraph beginning with the name of a researcher. Instead, **organise the literature review** into sections that present themes or identify trends, including relevant theories.



Figure 23: Application of the modality principle of Cognitive Theory of Multimedia Learning in design condition 2 (bottom) as compared to design condition 1 (top)



Figure 22: Application of the segmenting principle of Cognitive Theory of Multimedia Learning in design condition 2 (bottom) as compared to design condition 1 (top)

The classical scientific method - problems of inductivism

Can we actually make a universal claim based on a finite number of observations?

In science it is belief that theories can never be proved, but only disproved. There is always a possibility that a new observation or experiment will conflict with a long-standing theory.

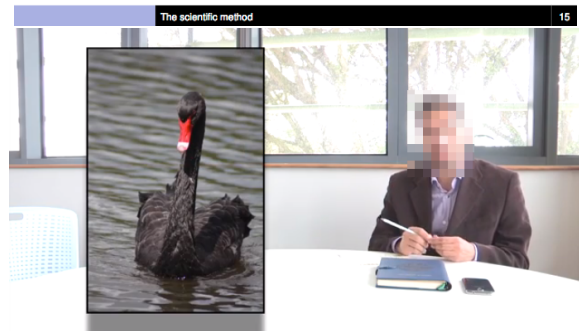
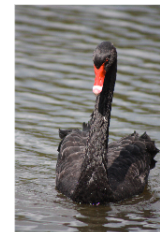
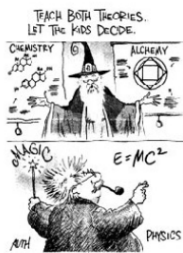


Figure 24: Application of the multimedia principle of Cognitive Theory of Multimedia Learning in design condition 2 (bottom) as compared to design condition 1 (top)

Pseudo Science



Pseudoscience is a claim, belief or practice which is **incorrectly presented as scientific**, it does not adhere to a valid scientific method, it cannot be reliably tested, or otherwise **lacks scientific status**.⁶

⁶Oxford English Dictionary



Figure 25: Application of the personalisation principle of Cognitive Theory of Multimedia Learning in design condition 2 (bottom) as compared to design condition 1 (top)



Figure 26: Application of the image principle of Cognitive Theory of Multimedia Learning in design condition 2 (bottom) as compared to design condition 1 (top)

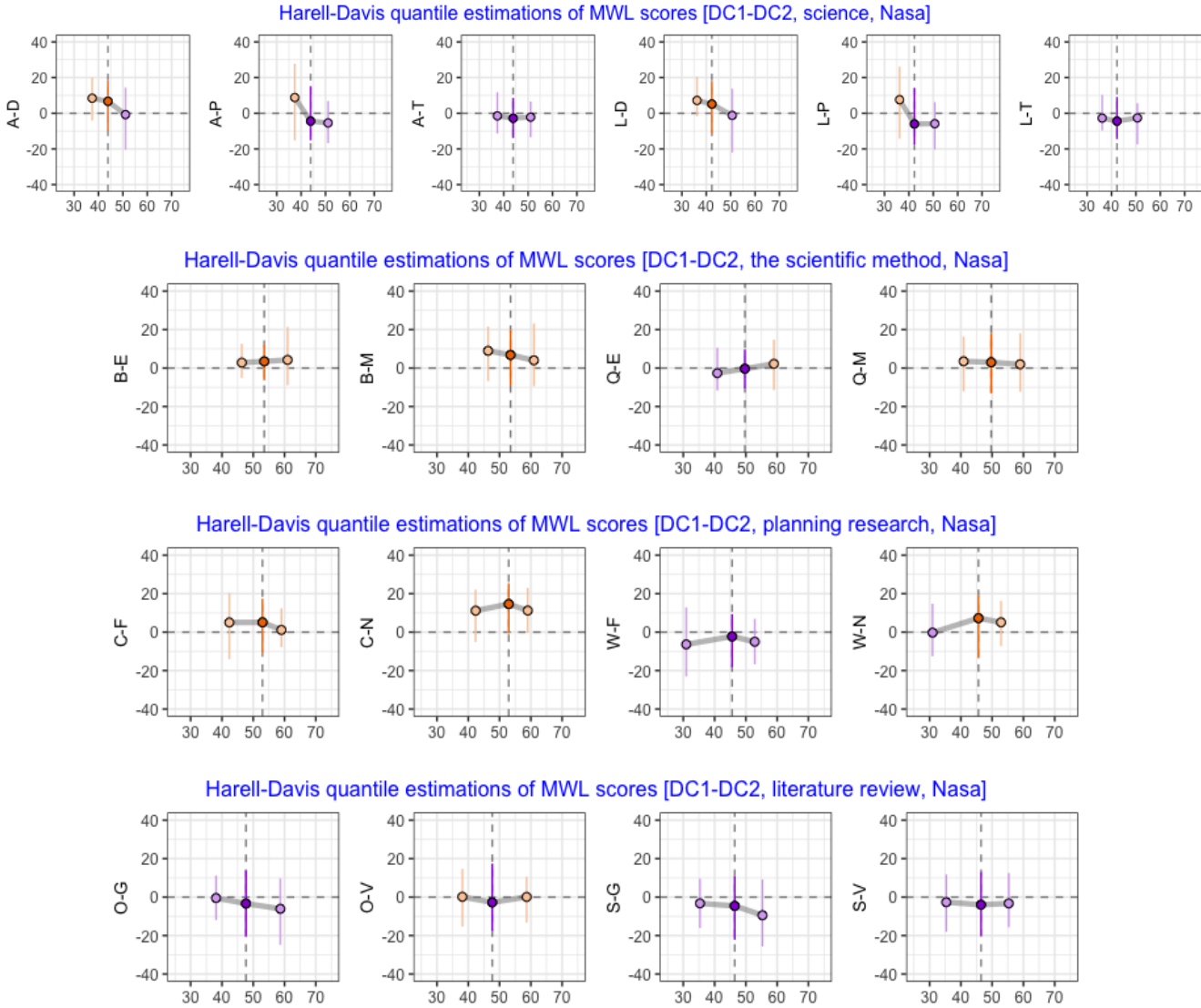


Figure 27: Harrell-Davis quantile estimations of the mental workload scores between design condition 1 and 2 for the Nasa Task Load Index instrument, grouped by taught topic. Horizontal lines and dots respectively represent the confidence interval and the mean for that quartile. Yellow and violet lines indicate the higher mental workload scores respectively for DC1 and DC2.

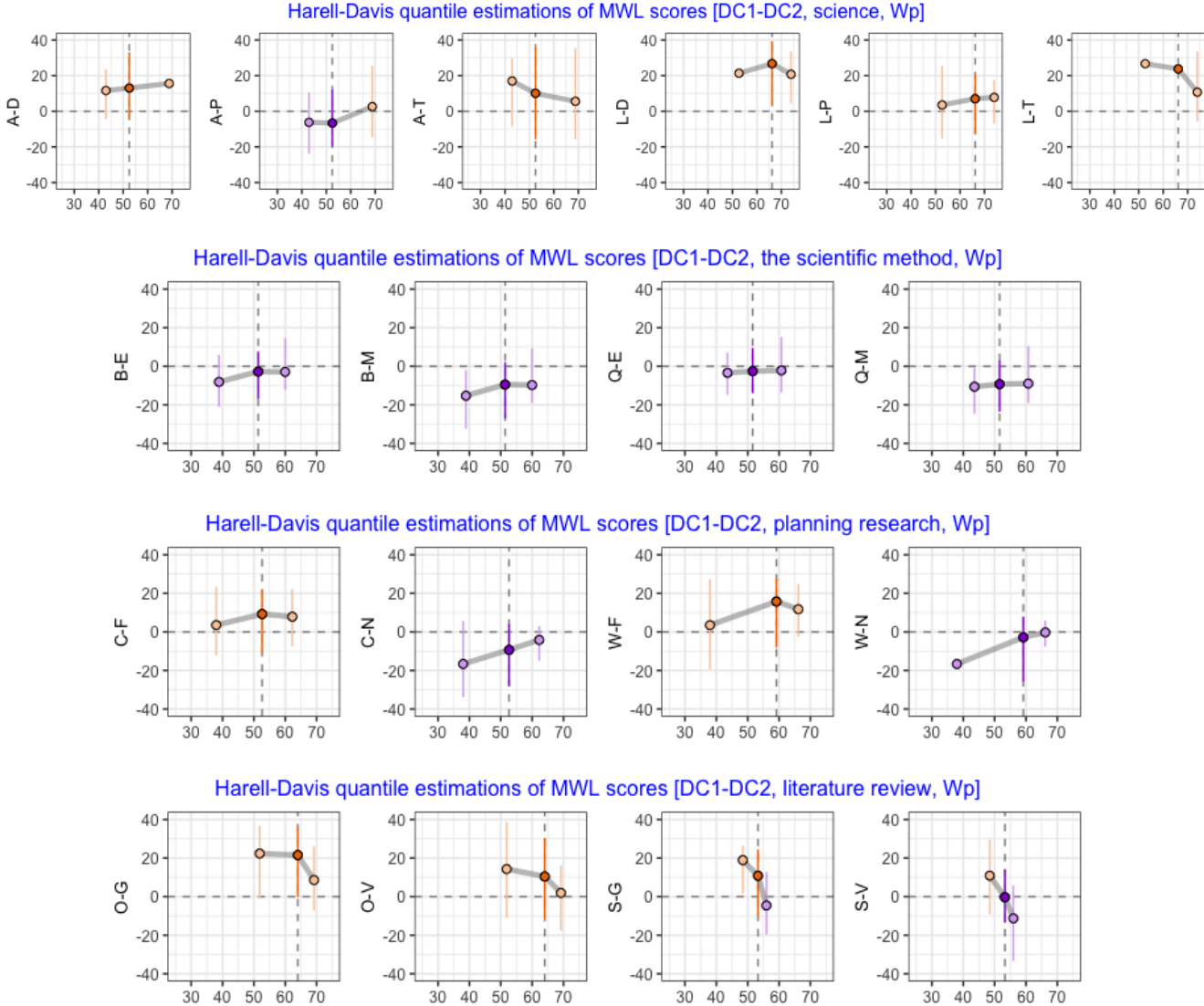


Figure 28: Harrell-Davis quantile estimations of the mental workload scores between design condition 1 and 2 for the Workload Profile instrument, grouped by taught topic. Horizontal lines and dots respectively represent the confidence interval and the mean for that quantile. Yellow and violet lines indicate the higher mental workload scores respectively for DC1 and DC2.

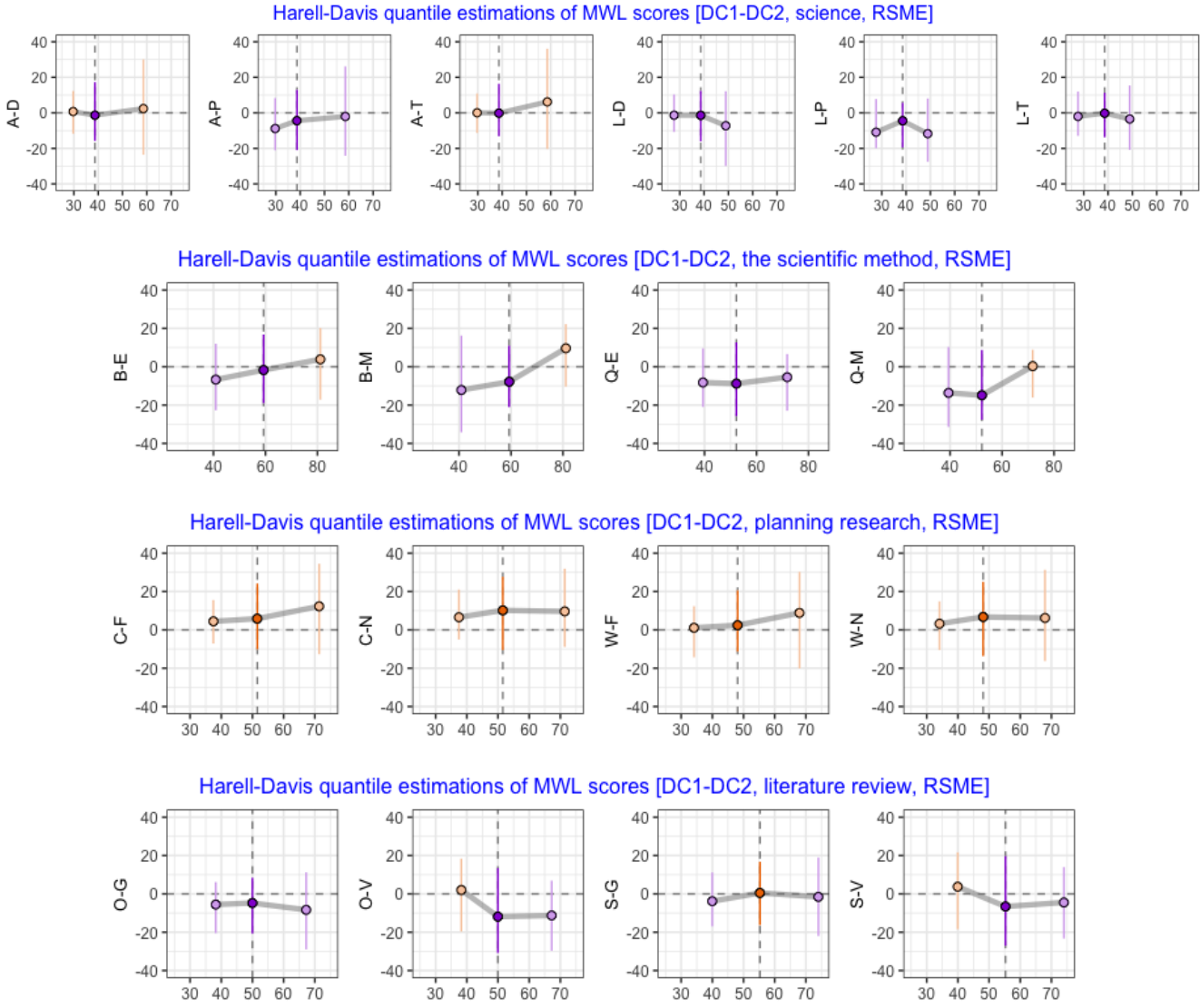


Figure 29: Harrell-Davis quantile estimations of the mental workload scores between design condition 1 and 2 for the Rating Scale Mental Effort instrument, grouped by taught topic. Horizontal lines and dots respectively represent the confidence interval and the mean for that quartile. Yellow and violet lines indicate the higher mental workload scores respectively for DC1 and DC2.