

Measuring cognitive load and cognition: metrics for technology enhanced learning

Stewart Martin

Abstract

This critical and reflective literature review examines international research published over the last decade to summarise the different kinds of measures that have been used to explore cognitive load and critiques the strengths and limitations of those focussed on the development of direct empirical approaches. Over the last 40 years, cognitive load theory has become established as one of the most successful and influential theoretical explanations of cognitive processing during learning. Despite this success, attempts to obtain direct objective measures of the theory's central theoretical construct – cognitive load – have proved elusive. This obstacle represents the most significant outstanding challenge for successfully embedding the theoretical and experimental work on cognitive load in empirical data from authentic learning situations. Progress to date on the theoretical and practical approaches to cognitive load are discussed along with the influences of individual differences on cognitive load in order to assess the prospects for the development and application of direct empirical measures of cognitive load especially in technology-rich contexts.

Keywords: cognitive load; working memory; measurement; multimedia; technology; learning; learning analytics

To reference this paper: Martin, S. (2015). Measuring cognitive load and cognition: metrics for technology-enhanced learning. *Educational Research and Evaluation*, 20(7-8), 592-621.

Introduction

One of the defining characteristics of educational provision in recent years has been the increasing interest in collaborative and blended e-learning, fuelled by the increasing ubiquity of educational technology and social media and their many affordances. Pressures for the use of e-learning and interest in more collaborative pedagogies have also arisen from desires to enhance institutional profile and “outreach”, from student demand, the prevalence of open-access resources and interest in moving away from a top-down instructor-led pedagogy.

These pressures have so far been felt more in higher and further education than in mainstream provision but have combined to raise questions about the role of the traditional instructor, about how to promote greater student engagement, and for ways to satisfy student demands for learner-centred provision and respond to changes in funding regimes that have increasingly elevated the importance of the student voice across all educational sectors.

Such debates raise questions about the effectiveness of e-learning and collaborative learning

and interest in how we might develop blended learning and an instructional model that is able to guide developments of technology use. Current technologies offer attractive possibilities for more effective learning, may encourage anywhere anytime education, may improve social interaction and enable individualised learning, efficiency gains for institutions, and stronger student engagement (Shuler, 2009). Advocates of technology use have argued that contemporary students are more attuned to learning with technology – that they even think and learn differently than previous generations of students. Although technology is felt to offer the prospect of more effective learning, there is relatively little empirical evidence showing that gains in educationally desirable outcomes are greater with the use of technology than with traditional approaches to learning. This is of special concern where content retention may be becoming a necessary but insufficient output of education in a world where higher order thinking skills, creativity, and problem solving are often seen as the most valuable cognitive attributes for future citizens and the industries and economies that will sustain them. So, whilst the growing use of digital technology and its application to learning is argued to have the potential to significantly improve instructional efficacy, particularly with regard to the successful learning of information and the development of understanding or skill (Mayer, 2008; Miller, Chang, Wang, Beier, & Klisch, 2011), concerns persist about the degree to which the design and use of digital learning materials have realised or optimised such potential (Argyris, 1976; Massa & Mayer, 2006; Schnotz & Kürschner, 2007; Sweller & Chandler, 1994; Tabbers, Martens, & Van Merriënboer, 2000).

To ensure that educational interventions in e-learning, collaborative learning, and also more traditional approaches to learning are optimum and effective, there is a need for instruments with which to assess the outcomes from such learning in complex and ill-structured domains as well as those of factual knowledge. Such assessments require a theoretical model of learning that is accessible to empirical verification and extends beyond mapping the retention of information, and for these reasons the concept of cognitive load and cognitive load theory have attracted much interest.

Although there remains a preponderance of single-study work in the field of cognitive load research, there has been activity across a wide range of subject areas from mathematics, reading, physical exercise, repetitive spaced learning and testing, and in a range of approaches including “brain training” learning games, embodied cognition (learning through action), and the personalisation of experiential learning. Pedagogical approaches have been similarly varied although the evidence of improved engagement and increased learning with learning games is as yet limited to results from young adults (Howard-Jones, 2014), and the established perceptions of educators about the efficacy of particular methods are often not supported by research evidence (Coe, Cesare, Higgins, & Major, 2014; Martin, 2010).

This systematic review provides a comprehensive introduction to cognitive load and its measurement and a synthesis and critical review of research findings in the field particularly over the last decade for both newer and experienced researchers. This review also identifies how previous research has approached the measuring of cognitive load in relation to learning outcomes and summarises, evaluates, and critically reflects upon the implications.

Context

The application of cognitive load theory (Sweller, Van Merriënboer, & Paas, 1998) has been at the forefront of much experimental work in cognition and has important implications for optimising the design of educational multimedia (Martin, 2012; Mayer, 2003, 2009; Mayer & Moreno, 2002). However, some writers express concern that the theory has developed little since its proposition in 1998 (Moreno & Park, 2010), whilst others note that despite its great influence on educational research in recent years cognitive load theory is by its formulation difficult to disprove; a difficulty compounded by the lack of means by which cognitive load may be measured directly (De Jong, 2010).

From Karl Popper's perspective of critical rationalism, cognitive load theory cannot be regarded as a truly scientific theory because several of its fundamental assumptions are not falsifiable, as they cannot be tested empirically (Popper, 1959, 1963). For each of the three types of cognitive load proposed by the theory (germane, extraneous, and intrinsic cognitive load), similar assumptions are made: For example, with regard to extraneous cognitive load, this is assumed to be caused by poor instructional design; and it is also assumed that cognitive processes that do not support schema construction or automation result in extraneous cognitive load; and that this is harmful for learning.

The difficulty with adhering to strict Popperian concerns about falsifiability is that it leads to a situation where hardly any theory could be regarded as scientific. Many well established scientific or logical statements and theories are based on fundamental assumptions that are presupposed by the methods used to test them and are not therefore falsifiable (see the "some swans are white" discussion in Blanshard, 1962). We tend to accept and apply many ideas that cannot be proved to be true but provisionally accept them because they can usefully be discarded if they can be proved to be false. So, we tend to use theories in science primarily because they are able to explain our observations of the empirical world and we replace them when theories with even more explanatory power appear, as happened when Darwin's theory of evolution by natural selection replaced the Lamarckian theory of evolution by the inheritance of acquired characteristics.

So, critical rationalists would regard it as improper for researchers to assume, for example, that any increase in measured cognitive load would result in poorer learning (or vice versa), as this presupposes the assumption that higher cognitive load is unhelpful for learning (an unproven hypothesis). It would therefore be necessary to remove such presuppositions about the outcomes of experiments that are embedded within the rationale for the experiment. To address such concerns, cognitive load theory would also need to use measurement instruments that were able to accurately discriminate between the three types of cognitive load. This remains problematic in much published research, where many established instruments are not sensitive to different types of cognitive load and provide only one composite measure of cognitive load (see Paas & Van Merriënboer, 1994, for physiological measures; Brünken, Plass, & Leutner, 2003, and Callan, Sutton, & Dovale, 2010, for dual-task measures; and Paas, 1992, for single-item rating scales). However, it may be possible for more recently developed objective measures such as functional near infrared spectroscopy (fNIRS) to help meet Popperian concerns without invoking the assumptions that provoke the criticism of structural rationalists (Cierniak,

Scheiter & Gerjets, 2009).

The review

Previously conceived as “mental load” (Moray, 1979), research on cognitive load has been of significant interest for over forty years, during which time it has developed into what is now a leading theory for describing cognition in learning, especially with digital technologies (Chandler & Sweller, 1991; Martin, 2012; Mayer, 2001; Niegemann, 2001). The concept of cognitive load and its associated theoretical framework are increasingly used to guide the instructional design and evaluation of educational multimedia environments and web-based instruction. However, despite decades of research on cognitive load, the direct measurement of its key concept – cognitive load itself – remains elusive. Much research on learning with digital technologies routinely uses cognitive load to explain individual differences in learning outcomes but does not directly measure the actual cognitive load experienced by learners (Brünken, Steinbacher, Plass, & Leutner, 2002). This is not because of any lack of interest in obtaining such measures from the research community but is attributable to a number of deep-rooted practical and theoretical issues that make achieving reliable and valid measures of cognitive load challenging.

As a result of its widespread adoption, research and interest in cognitive load theory ranges across many disciplines and contexts, and locating relevant scholarly studies therefore involves searching a wide range of sources for appropriate material. The focus of the present study was on empirical work where the measurement of cognitive load or critiques of such measurement or of the instrumentation used or available to conduct such measurements was a significant part of the work. The use of a systematic review was therefore selected as this method uses a well-defined and rigorous approach that attempts to identify, appraise, and synthesise all the empirical evidence that meets pre-specified eligibility criteria to answer a given research question. The use of this explicit method aims at minimising bias, in order to produce more reliable findings that can be used to inform decision making and future research (Davies, 2000; Gough, Oliver, & Thomas, 2013; Higgins & Green, 2011).

Key steps in a systematic review include the identification of relevant studies from a number of different sources, selecting studies for inclusion and conducting an evaluation of their strengths and limitations on the basis of clear, predefined criteria followed by the systematic collection and appropriate synthesis of data. The research question was Under what circumstances and with what instruments can cognitive load be measured in valid and reliable ways? The minimum selection requirements for inclusion of material were therefore about fitness for purpose and whether the material included empirical work and the conducting or critique of measurements of cognitive load or of the instrumentation used to conduct such measures. For potential inclusion, sources also had to include content on measuring cognitive load in reliable and valid ways, ideally that are therefore potentially transferrable across knowledge domains. The review was also on locating relatively recent academic articles that were substantially about instruments or techniques that reported some success in achieving this, particularly if they gave detail or analysis of the metrics obtained, as discussed above. Because of the subject range of cognitive load theory in the research literature, a total of 52 major academic databases were therefore searched

for appropriate material (Table 1).

Table 1. Databases searched.

Brill Journal Archive	LexisLibrary
Cambridge journals online	Libraries of the world
CERUK	Literature online
Clinical Evidence	Medline
Cochrane library	MIT Press Journals
COPAC (including British Library integrated catalogue)	NBER working papers
Digimap	NDLTD Networked digital library of theses and dissertations
EBook Library (EBL)	New England journal of medicine
EBSCO (including Business source premier, Cinahl Plus, ERIC, PsycARTICLES, PsycBOOKS, PsycINFO and SPORTDiscus)	NFER
Embase	OECD-iLibrary
Emerald journals	Oxford Journals online – includes Oxford journals archive
FAME	Project Muse
Harvard University library digital projects	Prola
IEEE Xplore	Proquest (including ERIC and ProQuest Dissertations and Theses)
Index to theses (ASLIB)	PubMed
Informa Healthcare	SAGE Premier (including SAGE Archives)
InfoSci-journals	Science Direct
IngentaConnect	Scientific World
Institute of Physics Press	SCIRUS
International Institute for Asian Studies (IIAS)	Scopus
INTUTE	Springer LINK
JSTOR	St. Jerome Publishing
Lancet	Swetwise
	TechXtra

The search variables employed were cognitive load and all derivatives of “measure” including measure, measuring, measurement, measures, and so forth. As the review was mainly focussed on activity over the last decade, the search included only items where the full text was available online and that were published from January 2004 onwards. The search located a total of 10,796 items including books or eBooks (415), book chapters (124), conference proceedings (1,008), dissertations and theses (234), journal articles (8,719), papers (1), reference texts (8), and reports (11). Material from unpublished dissertations, conference papers, or technical reports was not included as their selective availability seemed likely to bias the systematic review.

Two-hundred-and-seventy-six items were excluded on the grounds of academic unsuitability or quality, for failing to meet the minimum requirements for inclusion, and/or where there was little or no information about the instrumentation that was used; these tended to be “overview” or promotional pieces with relatively little detail and comprised book reviews (42), magazine articles (66), newsletters (53), newspaper articles (114), and one trade publication. The remaining 10,520 relevant sources were then filtered for scholarly peer-reviewed publications only where academic critique was brought to bear on any claims made for validity, reliability, or the replicability of the outcomes that were claimed for the

measurement approaches described, which reduced the total of relevant items to 7,097 articles. Each of these shortlisted articles was reviewed manually for suitability of focus within title, abstract and/or keywords, and concluding sections, and this process identified 136 articles, each of which was read in its entirety.

Current models of cognitive load theory are based upon the earlier work of schema theory (Anderson, 1983), dual-coding theory (Paivio, 1986), working memory models (Baddeley, 1986), and generative theory (Wittrock, 1974). Early measures of cognitive load emerged from item difficulty tests (Bratfisch, Borg, & Dornic, 1972), but these were limited by their subjectivity (Paas & Van Merriënboer, 1994). This earlier work was subsequently integrated into a theoretical framework developed by Sweller (1999) and incorporated within Mayer's theory of multimedia learning (Mayer, 2001).

Cognitive load theory seeks to explain how and why some material is more difficult to learn than other material and is based on the proposition that the human brain uses two types of memory: short-term (working) and long-term (storage) memory, where short-term memory is seen as having a limited capacity, perhaps for as few as four "chunks" of information (Halford, Baker, McCredden, & Bain, 2005) and long-term memory is seen as having almost unlimited capacity (Sweller, 1994). Working memory is commonly defined as "a brain system that provides temporary storage and manipulation of the information necessary for such complex cognitive tasks as language comprehension, learning, and reasoning" (Baddeley, 1992, p. 556). Working memory therefore represents a limit on learning in terms of storage and working space for conscious cognition. Working memory is thought by some to be fixed genetically, to grow with age (Miller, 1956) and to limit our "thinking-holding space" (Johnstone, 1997) when processing or holding information because the more information that has to be held the less space there is for processing. In terms of developing greater understanding, learning therefore ceases when working memory becomes overwhelmed by too much information and processing (Reid, 2008). Cowan (2000) argues from an extensive review of research and theory that different underlying forms of memory representation all indicate similar capacity estimates for working memory. Although at any one time different memory chunks may be less or more prominent in memory representation, the focus of our attention will determine how many of the most prominent chunks in the representation we can attend to at once, and four chunks appears to be a functional limit for optimal processing.

The existence of long- and short-term types of memory in humans is thought to be important because it significantly influences the way we learn. Using short-term memory, we develop schemas ("cognitive constructs that incorporate multiple elements of information into a single element with a specific function" – Paas, Renkel, & Sweller, 2003, p. 2) and store these in long-term memory. Schemas help us when solving problems that we have not seen before by drawing on our learning about similar kinds of problems we have solved in the past and thereby speeding up problem solving and task execution by partially automating our cognitive activity when responding to situations or problems that are similar to ones we have learned about in the past "by chunking individual elements into a single element" (Sweller, 1994, p. 299). We use the limited capacity of short-term memory to manipulate existing schemas (or to create new ones) and apply these to solving problems that would otherwise prove too complex for us to deal with if we always had to begin from

first principles.

Baddeley's widely used model of working memory assumes that a central executive exists in the human brain and that this coordinates two other interconnected but separate and independent systems, one auditory system that processes information such as music or spoken material – the phonological loop – and a second system that processes written material or pictures – the visuospatial sketchpad (Baddeley, 1986; Baddeley & Logie, 1999; Repovš & Baddeley, 2006). This dual-coding assumption was originally taken from dual-coding theory (Paivio, 1986). The phonological loop and the visuospatial sketchpad are thought of as slave systems and are presumed to have limited capacity and by virtue of being independent from each other are unable to compensate for lack of capacity in the other. If as a result of a learning experience the processing ability of either the visuospatial sketchpad or the phonological loop approaches zero, the learner experiences high cognitive load. The amount of this difference between the total cognitive load and the processing capacity of the auditory or visual system is seen as the free cognitive resource of the learner, and this has also been used to create one direct measure of cognitive load (Brünken et al., 2003). In an alternative proposal, Barrouillet and Verbal proposed a resource-sharing model for working memory which includes a time-based feature that assumes that rapid switching occurs during processing (Barrouillet & Verbal, 2004; Barrouillet, Bernardin, Portrat, Vergauwe, & Camos, 2007). According to the time-based resource-sharing model, the cognitive load generated by any given task is a function of the proportion of time during which it captures attention, during which it impedes other attention-demanding processes, and information in working memory is therefore likely to decay rapidly as soon as attention is captured by another activity (Barrouillet et al., 2007).

These and other theoretical models of the workings of the phonological loop and the visuospatial sketchpad have been difficult to research, not least because the actual amount of cognitive load that is created within these systems has proved difficult to measure directly. Further, some have suggested that individuals with higher visuospatial working memory capacity are affected differently by verbal working memory load than those with lower capacity, suggesting that capacity in one may affect that in the other (Ross et al., 2014). However, research on cognitive load routinely uses this theoretical rationale to explain different learning outcomes for individuals using multimedia or webbased learning resources (e.g., Brünken et al., 2003).

Cognitive load theory argues that whether particular material is easy or difficult to learn depends in large part on the degree to which we are able to manage the amount of processing (cognitive load) needed to solve a problem or learn something new by using schema acquisition and automation. Cognitive load theory proposes that three different kinds of cognitive load operate. Extraneous cognitive load is the difficulty associated with the design of instructional material, especially the way information is presented to the learner. High extraneous cognitive load inhibits learning because of unnecessary processing caused by the instructional design. Germane cognitive load is the load created by constructing, processing, and automating schemas (and can also be manipulated by the instructional design) but is helpful to learning because it results from features of the design that direct attention towards relevant learning processes. Intrinsic cognitive load is attributable to the inherent complexity or difficulty of the material to be learned; this cannot be

changed by the teacher, is assumed to be unaffected by the instructional design, and is thought to be the product of a combination of the learner's prior knowledge and the intrinsic complexity of the learning material (Sweller & Chandler, 1994).

Cognitive load theory proposes a mechanism whereby learner engagement in cognitive tasks can be understood in neurophysiological terms, with some suggesting that the core mechanisms of working memory are settled as early as during childhood (see Portrat, Camos, & Barrouillet, 2009). However, overall mental performance is a multidimensional construct, and cognitive load theory does not take account of factors such as the influence of individual goals, expectations, or beliefs on cognitive performance (Moreno, 2006; Moreno & Park, 2010). To achieve this would require a strengthening of the theoretical underpinning of cognitive load theory to link it more closely to such individual differences and to take individual motivation and development of expertise during courses of study or training into account and to flexibly adapt instruction by better assessment of a learner's expertise on the basis of their performance (Van Merriënboer & Sweller, 2005).

Researchers in the field of cognitive load theory seek to arrange the instructional control of cognitive load so as to optimise the load experienced by subjects in learning situations to avoid extreme situations where there is too little load or too much load, because learning deteriorates in both situations (Ayres & van Gog, 2009; Young & Stanton, 2002). Cognitive load researchers wish to produce both the optimum amount of load for learning and to promote load of the most helpful sort; that is, they seek to optimise the load that contributes to learning (i.e., germane load) and reduce the load imposed by elements that hamper learning (i.e., extraneous load). It is important to be able to separate and measure the three types of cognitive load because each of them is related to learning in different ways, but studies have found it difficult to empirically distinguish between their separate effects in a learning context (Brünken, Seufert, & Paas, 2010).

For example, when learners find instructional tasks easy (where intrinsic load is low), any extraneous cognitive load imposed by the learning resources or context may have little or no significant negative effect on learning. This is not the case when tasks are more difficult and the intrinsic cognitive load is high, and under these circumstances it is important to take appropriate account of the extraneous load on learners (Van Merriënboer & Sweller, 2005). However, in some learning situations it may be difficult to reduce the intrinsic load because the learning tasks may be unavoidably complex, they may have irreducible high element interactivity, or may require the use of many different schemas, such as in situations where multiple choices are available to the learner regarding the information to be selected and applied. Research using cognitive load theory has therefore sought to find ways to manage high intrinsic cognitive load (Pollock, Chandler, & Sweller, 2002) by approaches that measure and compensate for learner's prior knowledge (see Kalyuga, Ayres, Chandler, & Sweller, 2003) or that allow for the level of germane load imposed on learners by different instructional materials (Cierniak et al., 2009; Kalyuga, Chandler, & Sweller, 1998; Salomon, 1984).

As part of this approach, cognitive load theory has argued that the physical integration of multiple sources of information is generally beneficial for learners. Physical integration happens when, for example, text and images are combined in multimedia applications or on

the page of a textbook so that each does not simply replicate the content contained in the other. Physical integration reduces the need for learners to split their attention between (for example) separate illustrations and text on a page or screen, otherwise the learner's attention may be divided unhelpfully between the separate elements, as they attempt to process each one individually and make cognitive associations between them. This split attention effect is regarded as unhelpful for learning because it increases extraneous load and so learning materials featuring split-attention may overwhelm working memory capacity (Chandler & Sweller, 1992; Eilam & Poyas, 2008; Sweller, 1994). Educationalists using or designing multimedia or other technology-enhanced or e-learning resources would therefore need to guard against inadvertently provoking the split-attention effect in any resources they offered to students.

However, subsequent studies have found that in any given subject domain, certain learning resources which are beneficial for less expert learners can be disadvantageous as learners become more expert (Kalyuga et al., 1998). In particular, the physical integration of information as a means to minimise the split-attention effect becomes less helpful to learners as their expertise grows and it becomes counter-productive for learning as expertise increases still further (Kalyuga et al., 1998). For more expert learners, the physical separation of information can be more advantageous than its integration, because they are likely to already possess the schema that the learning resources are attempting to promote in less experienced learners. As a result, learning resources may become subject, therefore, to an expertise-reversal effect (Schnotz, 2010). The expertise-reversal effect appears when more expert learners find it easier to handle complex instructional material but more difficult to learn from material that is designed to integrate separate elements in order to aid less experienced learners to construct appropriate mental representations (schemas). In such cases, experienced learners are confronted with instructional guidance that is redundant for them, and this can be difficult to ignore, thus increasing cognitive load and reducing the efficiency of their learning (Kalyuga et al., 2003). Intrinsic cognitive load can therefore only be meaningfully measured with reference to a particular level of expertise (Schnotz & Kürschner, 2007). This could pose particular problems for educators wishing to avoid the expertise-reversal effect if they did not have reliable information about the expertise level of each learner in any given subject domain.

Obtaining valid and reliable measures of individual cognitive load from learners can be problematic for other reasons also, not least because learners faced with a new topic or domain may find it hard to report whether any difficulty they experience is due more to the content or to the instructional design. In such circumstances, it may be impossible to reliably identify and disentangle the origins of extraneous and intrinsic cognitive load (Cierniak et al., 2009). As a result of either one or another type of cognitive load being higher or lower for different learners, the overall load for different learners may be very similar but we may not be able to know the source from which the load originates. Different measures of cognitive load should therefore not be assumed to measure overall cognitive load, but it may be possible to use them to measure different types of load. Different learning materials (or learner characteristics) may produce different patterns of results that also might depend upon the prior knowledge of the learner (Kalyuga & Sweller, 2005).

Because intrinsic load varies not just as a result of the inherent complexity of the learning

material but also with the expertise of the individual learner in a given subject or content area, establishing the intrinsic load for individual learners is important for maximising their learning. This is not straightforward because the measurement of intrinsic cognitive load often relies on subjective or objective measurement instruments that have significant structural limitations. One of the main problems with subjective instruments in general, for example, such as self-report questionnaires, is attributable to the difference between espoused theory and theory in use (Argyris, 1976), which is the difference between what individuals say they do and what they actually do. A learner might employ entirely different strategies (or experience entirely different difficulties) in practice from those they consistently report in good faith on questionnaires or during interview.

The value of using subjective perceptions of cognitive effort to measure cognitive load is based on the assumption that individuals are able to make reliable and valid estimates of the amount of cognitive effort they are expending in a given situation. This approach has the great benefit of simplicity but has the significant drawback that it usually involves singlepoint post-hoc assessments with questionable content validity. It is also difficult to be sure that retrospective assessments are accurate assessments of the cognitive load expended (as opposed to what might in fact be assessments of the difficulty of the task), or to which type of load an individual assessment relates, or which type of load created the perceived effort, or to know how the assessment can be related directly to learning. Additionally, subjective measures are frequently administered post-hoc and may therefore be mediated by faulty recollection or rationalisation, or during activity breaks which may disrupt the activity being investigated, and subjective methods have on these grounds been criticised as suffering from low evaluative bandwidth (Lin, Li, Wu, & Tang, 2013). Despite these reservations, some studies have argued in favour of self-rating instruments on the grounds that participants have been able to detect differences in task complexity and have suggested that subjective measures may therefore in practice be sensitive to intrinsic cognitive load, be highly reliable, and also have the benefit of being relatively unobtrusive (see Ayres, 2006).

This view is supported by some studies interested in quantifying and evaluating cognitive load in high frequency interaction scenarios, such as found in air-traffic control. In one such study where task difficulty was mediated using the number and incoming direction of aircraft, participants were judged to be able to discriminate between the three types of cognitive load with over 74% accuracy (Lin et al., 2013).

Limitations also exist for other instruments proposed for measuring cognitive load especially when this varies as a result of the learner's changing framework of reference and increased schema acquisition in response to progressive learning (i.e., as learner expertise increases), because the difficulties that are perceived by the learner and the associated degree of helpfulness of particular resources may be continuously changing as learning proceeds (Schnitz & Kürschner, 2007; Veenman, Prins, & Verheij, 2003). However, work comparing different approaches for establishing item difficulty levels has shown that individual learners do perform quite well when judging the difficulty level of items and are likely to be more accurate than subject experts at establishing the "true" level of difficulty; that is, the level established from a large sample size (Wauters, Desmet, & Van Den Noortgate, 2012). Using student evaluations of item difficulty levels to help guide a determination of the subject expertise of a particular learner may therefore in practice be an

effective way to match different instructional material to a learner's developing mastery of content.

Measurement of cognitive load In the early stages of the development of cognitive load theory, measurement was undertaken mainly by indirect methods such as learner error rates, the time taken to reach problem solutions, or computational models that were often used alongside more established approaches such as dual-task methodology (Dutke & Rinck, 2006; Sweller, 1988).

These approaches have been largely superseded by self-rating Likert-scale measures where learners are asked to report the perceived amount of mental effort they invest in a learning experience (Paas, 1992), where mental effort is defined as "the aspect of cognitive load that refers to the cognitive capacity that is actually allocated to accommodate the demands imposed by the task; thus, it can be considered to reflect the actual cognitive load" (Paas, Tuovinen, Tabbers, & Van Gerven, 2003, p. 64). In order to explore more fully the different cognitive process within such global measures, a set of separate (but still subjective) scale items has been proposed to measure intrinsic, extraneous, and germane cognitive load individually (Cierniak et al., 2009).

Other measures have sought to assess the efficiency (quality) of learning by mapping the relationship between the mental effort invested and test performance (Paas & Van Merriënboer, 1993) or between the time invested and test performance (Gerjets, Scheiter, & Cierniak, 2009; Van Gog & Paas, 2008). Although still widely used, self-rating Likertscale measures have also attracted criticism on the grounds that they may be too subjective (Brünken et al., 2003), or because they have been too often used to measure different types of cognitive load inappropriately (Kirschner, Ayres, & Chandler, 2011), and because they have not always been used in a consistent way (Van Gog & Paas, 2008). Some writers argue that these different measurement approaches towards cognitive load have actually been assessing different constructs and therefore are not equivalent or necessarily comparable (e.g., Ayres & Paas, 2012; Ayres & Van Gog, 2009).

For the purposes of this review, the most commonly used mental workload metrics have been divided into the three broad but somewhat overlapping categories of subjective measures, physiological measures (direct objective), and secondary-task (indirect objective) approaches. However, the specific techniques used in each case may be accessing slightly different features of mental workload, and in practice there is thought to be some value in using multiple measures, preferably from two or more of the broad categories shown in Table 2 (Carswell, 2005).

Subjective measures

Subjective approaches most commonly use participant ratings of the difficulty of the materials to be learnt as a measure of cognitive load (Kalyuga, Chandler, & Sweller, 1998), although differences in ratings between participants can also be influenced by the difficulty of the task, the differing expertise or competency of individual participants, or other factors such as variations in participants' motivational levels or in their ability to concentrate.

Indirect subjective measures include self-reports of stress or the mental effort expended and commonly rely on the retrospective use of Likert-type scales or on data collected during task performance (e.g., Antonenko & Neiderhauser, 2010; Ayres, 2006; Chang & Yang, 2010; Kalyuga & Sweller, 2005; Scharfenberg & Bogner, 2013). However, it is not clear how these self-reports can be related directly to the actual cognitive load involved (Brünken et al., 2003). This issue is problematic for many other approaches to measuring cognitive load, where a small amount of mental effort could be taken to mean that the learning task generated a small cognitive load (it was an easy task), or that the task was difficult but the learner possessed high expertise (the learner found the task easy because of their high degree of competence), or it could equally indicate that the cognitive load demanded was so high that the learner gave up trying to understand or complete the task (irrespective of their level of expertise). These “efficiency” measures are problematic when some individuals reach similar levels of performance than others with less mental effort or where higher levels of performance than others are attained with the same amount of invested effort. In each case, the relationship between self-report values and the experienced cognitive load means something different and should be interpreted differently.

Correlations between self-report scales, secondary visual monitoring tasks, and post-hoc difficulty rating scales are not reliable or consistent and tend to be weak, and different types of cognitive load are often dissociated from each other, are not highly correlated, and can be sensitive to different measures of cognitive load, such as reaction times to secondary tasks, effort ratings during learning, and difficulty ratings after learning (DeLeeuw & Mayer, 2008).

If cognitive load is influenced by or composed of different elements as suggested (Mayer, 2001; Sweller, 1999), different manipulations in the learning situation would seem likely to cause different types of cognitive load to vary. It may therefore be dangerous to assume that different measures of cognitive load each measure overall cognitive load, but it may be that these may be used to measure different types of load. Different learning materials (or learner characteristics) may also produce different patterns of results which might depend on the prior knowledge of the learner (Kalyuga & Sweller, 2005).

These and similar complexities undermine any assumed linear relationship between performance and mental effort, especially when we additionally allow for differences between individuals with regard to interest in a given learning topic or levels of motivation, as these also seem likely to influence an individual’s investment of mental effort. Higher task demands are not necessarily associated with a higher mental workload because, as we have seen, mental workload cannot be precisely measured using only the properties of the task itself. Other, individual, factors such as expertise and environment have an impact on the mental effort that subjects deploy to solve a particular task (Ayaz, Cakir, et al., 2012; Ayaz, Shewokis, et al., 2012; Murai, Hayashi, Okazaki, Stone, & Mitomo, 2008). Cognitive load may therefore need to be defined in terms of the interaction between the given task and the individual involved in performing the task (Durantin, Gagnon, Temblay, & Dehais, 2014), and, despite years of research, it remains unclear how mental effort is related to actual cognitive load.

Table 2. Methods of measuring cognitive load (after Kalyuga, 2009; and Brünken, Seufert, & Paas, 2010).

Relationship	Data	
	Subjective	Objective
Direct	Self-reports of stress	Eye-tracking, Dual-task methods, Brain activity measures (e.g., MRI, fNIRS).
Indirect	Self-reports of mental effort	Learning outcomes, time on task, help-seeking behaviour, physiological approaches using EEG or cardiovascular metrics, efficiency measures.

Simple subjective rating scales continue to be the most common approach to representing cognitive load. Their reliability and validity rests on the questionable assumption that learners are able to reflect and report on their own cognitive processes accurately (Paas & Van Merriënboer, 1994), but by their very nature they are not able to provide us with an absolute scale for ratings of mental effort or cognitive load, although they can be useful for repeated comparisons of load with the same group of learners. Educators should treat overall measures of cognitive load with some care, and differentiation between germane, extraneous, and intrinsic load remains challenging for all current instruments. Different instructional materials may also produce different outcomes depending on the prior knowledge and individual characteristics of each learner, and it may be that cognitive load is an expression of the interaction of a number of these individual elements.

Direct objective measures

Direct objective measures include examples of eye-tracking, dual-task methodologies or brain-activity measures using neuroimaging approaches, such as functional near-infrared spectroscopy (fNIRS). The latter, whilst perhaps currently the most promising, is as yet inconclusive because the connection between cognitive (memory) load and prefrontal cortex activity is still imperfectly understood.

Dual-task methodologies (sometimes referred to as the dual-task paradigm) can be regarded as objective in the sense that they use the assumption of a limited central processing system in the brain to argue that learner performance on a simple (secondary) task, such as reacting quickly to a separate event, maintaining a counting task, or being able to remember items seen on a separate computer screen, can be used to add cognitive load to the same working memory processing system as the primary task. This assumption allows researchers to claim that the elapsed time or successful completion of the secondary task can be used as a direct and individual measure of the cognitive load demanded by the primary task, although some have concluded that dual-task methodologies may be intrinsically limited for use with complex instructional designs where the identification of an appropriate secondary task may present inherently complex challenges that may be very difficult to overcome (Kirschner et al., 2011). Using reaction times to a secondary task as a proxy measure for cognitive load has also highlighted the importance and difficulty

of choosing an appropriate secondary-task design (Block, Hancock, & Zakay, 2010; Brünken et al., 2002; Schoor, Bannert, & Brünken, 2012). If the contiguous secondary task becomes too intrusive, it may at times become the primary task, but secondary tasks may not affect cognitive load when the task demand is low, because individuals would possess sufficient working memory to successfully complete both primary and secondary tasks. Additionally, the secondary task needs to align with the theoretical structure of the theory, so that if we accept that the two subsystems of working memory for phonological and visuo-spatial information are mostly independent, a visual secondary task may not correctly reflect cognitive load in the auditory subsystem and vice versa, especially when our understanding of how the cognitive load in these two subsystems affects overall cognitive load is uncertain.

This approach seems likely to become more problematic with repeated use if we consider that subject performance may change over time simply as a result of learning from practice at completing secondary tasks. Additionally, the secondary tasks involved must demand access to the same mental resources as the primary task, the performance measures need to be reliable and valid, and the secondary task has to be simple, so that it does not prevent simultaneous learning processes, but strong enough to potentially consume all of the relevant cognitive capacity. The duration of a secondary task also appears to influence activity more than its degree of difficulty (Towse & Hitch, 1995). Secondary-task methodology is therefore regarded as a sensitive and reliable technique under certain conditions but has drawbacks that may limit its wider use; it interferes with the primary task, and this effect is more pronounced with more complex tasks or when cognitive load is high; and it is unsuited to real-time or continuous measures (Lin et al., 2013).

Process-tracking methods such as concurrent or retrospective verbal reports from subjects have also been applied along with eye-tracking studies that trace fixation patterns, eye blink, and pupil dilation to evaluate cognitive load and map shifts in user attention (e.g., Chen & Epps, 2013; Irwin & Thomas, 2010; Klingner, Tversky, & Hanrahan, 2010; Kramer, 1991; Theeuwes & Belopolsky, 2010; Zheng & Cook, 2012). Supporters of these approaches promote their advantages in collecting data in real time and critique other approaches such as performance scoring (e.g., accuracy or reaction times) and subjective self-ratings as being too reliant on overt and discrete participant responses and as postprocessing measures. Studies have found that blink activity (latency and rate), pupil size, and fixation duration and rate all correlate significantly with different levels of demand on working memory, with blink latency, pupil size, and fixation duration progressively increasing in line with task difficulty and blink rate, fixation rate, and saccade speed and size decreasing similarly. The hypothesis offered by such work is that eye blink, pupillary response to tasks, and eye movement each reflect different but complementary information about cognitive activity because they are each controlled by different parts of the nervous system.

Some features of eye activity may therefore be useful approaches for discriminating between different levels of cognitive load (Jacob & Karn, 2003). A significant amount of research literature supports the view that pupil size varies as a function of experienced cognitive load, the level of drowsiness, and mood, feelings, or attitude, or affective state (Grandchamp, Braboszcz, & Delorme, 2014), although the results from some studies in

this area (e.g., Chen, Epps, Ruiz, & Chen, 2011) conflict with the findings of others (Greef, Lafeber, Oostendorp, & Lindenberg, 2009; Van Orden, Limbert, Makeig, & Jung, 2009).

The application of such research may still rely on self-report measures, where a number of oculometric measures such as blink frequency, pupils size, and gaze position may be used to monitor behaviour such as self-reported mind wandering, where participants use a button press to indicate occasions when they had forgotten to continue a counting exercise (Grandchamp et al., 2014). One limitation of this approach, as with other self-report systems, is that it still relies on the validity and accuracy of user reports. In the above study, for example, it is unclear whether subjects reporting fewer mind-wandering episodes than others could have experienced fewer of these or instead may simply have been less sensitive in detecting them. Other studies have suggested that the degree of pupil dilation can be affected by conditions such as depression, making this approach to measuring cognitive load ungeneralisable to the wider population (Siegle, Steinhauer, & Thase, 2004). Also, pupil dilation is regarded by some as an imperfect measure of brain activity because it does not control for many factors such as the brightness of the stimulus applied, or tiredness. Additionally, many eye-related studies of cognitive load make use of the Stroop Colour Naming Task, which is a widely used test of selective attention that requires participants to name the colour of the ink in which words that are the names of colours are written. Generally, people make more mistakes when the word is not the same as the colour in which it is written (e.g., the word “blue” printed in red ink) than when the word and the colour are the same (for a review, see MacLeod, 1992).

However, results from such studies are felt to be influenced by other factors such as participant attention, individuals reflecting on the previous stimulus, and also any mental processing that is irrelevant to the task about which, of course, the experimenters would have little or no information (Siegle et al., 2004). In particular, the use of the Stroop Test has been shown to create increased distractor interference effects when cognitive load becomes high (Gibbons & Stahl, 2010).

Although eye movement, pupil dilation, and blink rate are generally regarded as sensitive psychophysiological metrics for monitoring cognitive load (Klingner, Kumar, & Hanrahan, 2008; O’Brien, 2006; Palinko, Kun, Shyrovkov, & Heeman, 2010), they have at times been found unsuitable for this use in older people, where, for example, mean pupil dilation has been found to increase as a function of memory load in younger participants but has not been found to be sensitive to memory load for older participants (Van Gerven, Paas, Van Merriënboer, & Schmidt, 2004). The usefulness of this approach as a sensitive correlate for fluctuations in memory load may therefore diminish with age, and age-related differences in the sensitivity of the pupil to cognitive load would therefore make it difficult to compare different age groups because different levels of dilation may not correlate with the different levels of experienced cognitive load.

Brisson’s comparison study of pupil dilation using different eye-tracking equipment systems also found that each system created different errors and that the written construction and layout of the language being observed affected participant “arousal” as measured by the three different systems (Brisson et al., 2013). In this study, a language like

English (left-to-right, top-down sequence, and layout) could “arouse” participants most at the start of reading but less towards the end, irrespective of the subject or content of the text. Unless equipment from different manufacturers is able to compensate for point of gaze in pupil size measurement, the use of eye tracking methodologies will need careful experimenter calibration to ensure the reliability and replicability of their work.

So, although concurrent or retrospective verbal reporting, eye tracking, gaze fixation duration, and concept mapping may be useful measures of cognitive load, they may actually be measuring different constructs within this. As noted by Van Gog, Kester, Nievelstein, and Paas (2009), subjects who report less invested mental effort can also have higher mean fixation durations during parts of a task than others reporting higher invested mental effort, which may be because data on mental effort usually relate to overall task processes whereas mean fixation data are often calculated for parts of task processes. Concurrent and retrospective verbal reporting are also problematic because they result in different types of information and seem to be closely related to the time at which the reports are generated (Van Gog et al., 2009). It seems probable that they therefore rely on different memory systems and that different techniques such as verbal reporting, eye tracking, and concept mapping are sensitive to different mental processes or cerebral structures, and such concerns strengthen the argument for the use of combined measures approaches. However, asking participants questions or to explain their experiences tends to provoke reflection, which might lead to the construction of information that was not in fact part of the process. On the other hand, whilst concurrent reporting during instruction overcomes the risk of post-hoc reports being inaccurate, its application draws on working memory, and it may therefore become difficult to maintain under high cognitive load.

Other approaches have used computer-based studies to infer the experiences of learners from activity logs or real-time playback recordings to record cognitive-load experiences and the problems that learners meet. Direct objective measures are probably the only ones that could claim to measure cognitive load without having to rely on subjective information such as self-report data, indirect measures, or other (e.g., physiological) factors that may have only a tangential link to cognitive load and be potentially strongly affected by other factors, but these are by no means free of difficulty, and data from them when used in educational settings need to be interpreted cautiously.

Direct objective measures of cognitive load are highly desired by educators and educational assessment systems internationally, but instruments attempting to secure these require very careful use and interpretation by educators and many still rely on elements of self-report. Others have produced conflicting reports and, as with subjective measures, a number of studies have suggested that different constructs within overall cognitive load, or even other mental processes, may be being measured along with a range of individual physiological features.

Indirect objective measures

Indirect objective measures include physiological approaches using electroencephalography (EEG) or cardiovascular metrics or learning outcome measures, often combined with interaction features. These are regarded as objective because they are perceived as

measures of intellectual performance and as indirect because the factors measured are affected by the information processing and retrieval process that is taking place. There is a long history of the use of indirect objective approaches to evaluate stress, affective, and arousal states especially when psychological measures are correlated with physiological signals such as speech and linguistic features, facial expression, eye/body movement, or galvanic skin response (Shi, Choi, Ruiz, & Taib, 2007; Yap, Ambikairajah, Choi, & Chen, 2009; Yap, Epps, Ambikairajah, & Choi 2011). Many speech features have also been used to measure cognitive load such as speech rate, pause rate, onset delay, and interruption rate, along with mouse-speed and pressure, linguistic patterns, heart rate, performance methods such as error-rates or tests, and subjective methods of rating experienced loads (e.g., Khawaja, Ruis, & Chen, 2007; Lin et al., 2013; Yin, Chen, Ruiz, & Ambikairajah, 2008). Such measures have been found to be helpful because they are continuous and allow very fine granularity in the measure which can also be repeated at high rates, but for the most part they are probably outside the scope and resources available in the majority of educational settings at this time.

Physiological measures can include heart-rate monitoring, EEG measures, eye tracking, and pupillary response and assume that changes in cognitive load are highly sensitive to physiological changes which are equally assumed to be involuntary and therefore objective and have claimed high levels of classification accuracy, but it is recognised that such outputs are also sensitive to confounds of subject expertise, age, and physical or mental variations (e.g., Khawaja et al., 2007; Yin et al., 2008). For example, working memory span appears to develop as children grow older (e.g., to around age 8–9) such that older children can more easily switch their attention from their immediate mental task and can exert more control over their attention, perhaps as a result of developing faster and more efficient mental processing and more attentional capacity (Gavens & Barrouillet, 2004). Individuals with high- and low-working memory spans differ in their ability to suppress task-irrelevant behaviour and information (e.g., extraneous cognitive load). Low-span individuals appear to find it more difficult to maintain task goals in working memory, and this is argued to be due to their poorer attentional control (Unsworth, Schrock, & Engle, 2004). Low-span and high-span individuals differ in particular in the speed with which they can switch attention between tasks or sub-elements of tasks whilst learning, although they may in addition differ in their ability to process information, and both of these may affect results (Unsworth et al., 2004).

Studies using indirect objective measures often feature a number of groups of participants each using some variation of instruction about the same material. It is assumed that the intrinsic cognitive load imposed is the same in all cases as the material to be learnt is the same for all participants and, perhaps as a result of this feature, the use of this approach is more common in educational settings than others discussed here. The assumption is that the more effective and efficient the instructional method is, the more students will learn and the lower will be the extraneous cognitive load due to the instructional process. However, much can depend on the kind of measures that are used to test the outcomes, such as how long a learner takes to complete a task, and some studies have identified this as a serious weakness in this approach (Brünken, Steinbacher, Schnotz, & Leutner, 2001; Mayer, 2001), whilst others have noted that such learning outcomes are also consistently influenced by individual learner traits (Mayer, 2001; Plass, Chun, Mayer, & Leutner,

2003). Behavioural and physiological measures using eye-tracking, heart-rate, or pupildilation analysis similarly assume that changes in cognitive load are highly sensitive to physiological changes which are thought to be involuntary and therefore objective. However, each has the drawback of providing only indirect connections to cognitive load, and high levels of such activity may be due to individual differences in stress or emotional reactions to the learning materials, which may be unknown to the instructor in a given educational setting.

The left brain hemisphere is involved in speech and memory, whereas the right hemisphere is necessary for physical response to external events and for attention shifting, and electroencephalography (EEG) and magnetic resonance imaging (MRI) are two established technologies for measuring and classifying such brain activity, although care is needed in interpreting outputs from these and similar approaches (Freeman, Ahlfors, & Menon, 2009). EEG records electrical activity across the surface of the brain arising from currents within the neurons in the brain, and these data can be used to detect how long it takes the brain to process certain stimuli, such as those experienced during learning. During learning, the neurons in different active areas of the brain will consume more oxygen and therefore also generate different levels of a tiny magnetic signature which can be detected by the powerful magnetic fields generated within a MRI scanner. Volumetric images of these different signatures can then be created from the different levels of blood-oxygenation level-dependent contrasts to make a functional image (map) of the neural correlates of cognitive tasks.

Recent research in neuroscience has developed a number of instruments for studying brain function during problem solving and learning. Event-related functional magnetic resonance (fMRI) has been used to map the amount of hemodynamic activity (oxygen saturation and blood flow) in regions within the brain during learning. This provides researchers with longitudinal measures of rapidly changing cognitive activity across a range of specific brain areas, and, it has been argued, may even be able to measure the different types of cognitive load (Whelan, 2007). However, at present the bulky and sophisticated equipment needed make the use of fMRI make it impractical for all but the most specialised contexts, and this effectively rules it out as a practical application for almost all educational settings.

Grimes, Tan, Hudson, Shenoy, and Rao (2008) used EEG to measure visual task difficulty and extract related data for cognitive load, but could not easily distinguish between the different sub-types of cognitive load, and determining task difficulty proved difficult. As EEG measures are particularly sensitive to the specific user task, this was a significant limitation. In a later study, Anderson et al. (2011) used EEG to estimate and classify the cognitive effort dedicated to holding information in the mind for short periods of time while performing another cognitive task. The results showed high memory load classification accuracy both within and across different tasks, although variations across different users were substantial. Instead of classifying individuals according to cognitive load, studies such as these may therefore suggest that one potential way forwards might be to group together those who exhibit similar characteristics in their EEG signal and then apply different models to classify their load.

As an alternative to the bulky and highly expensive EEG and MRI/fMRI technologies, functional near infrared spectroscopy (fNIRS) was developed in the 1990s and observes closely similar physiological parameters to fMRI (Chance et al., 1998; Strangman, Culver, Thompson, & Boas, 2002). Because it can be used to measure activity in localised areas of the brain, most commonly in the anterior prefrontal cortex, it has emerged over the last decade as a promising technology for brain imaging. The approach takes advantage of the way light at near-infrared frequencies penetrates bone and biological tissue but is absorbed by haemoglobin in the bloodstream. Neural activity (such as learning or cognitive task completion) is accompanied by increased oxygen demands by neurons in order to metabolise glucose for energy, and fMRI detects the opening of capillaries in cortical areas where carbon dioxide has accumulated in response to neurons burning glucose.

First proposed in 2004 as a portable brain-computer interface suitable for long-term use, fNIRS uses near-infrared light to detect levels of oxygenated and deoxygenated haemoglobin on the surface of the prefrontal cortex (Coyle, Ward, Markham, & McDarby, 2004) and is a low-cost non-invasive neuroimaging technique that has been shown to be sensitive to both cognitive loads and states and is a viable alternative to fMRI (Fishburn, Norr, Medvedev, & Vaidya, 2014). The validity of fNIRS measurement has been repeatedly confirmed in recent years, and results can be reliably reproduced, even over time spans of 1 year. Studies also have found that fNIRS results are highly consistent with fMRI findings (Afergan et al., 2014), but its great advantage over MRI and EEG is that it has become increasingly portable with relatively low-cost wireless instruments now available, and because it uses much less bulky equipment it can more easily be used in authentic learning situations (Ferrari & Quaresima, 2012). Applications of fNIRS are steadily moving closer towards use in routine classroom situations, and this promising technology seems set to make inroads into mainstream educational research in the near future.

Some studies have used fNIRS to determine that haemodynamic levels in certain brain regions are related to problem-solver expertise, explaining at least 73% of the variation between participants solving tangram puzzles, suggesting that better problem solvers increasingly become less dependent on neuronal resources as they continue working on subsequent tangram puzzles (e.g., Cakir et al., 2011). Similarly Shaw, Satterfield, Ramirez, and Finomore (2013) used Doppler tomography during a communication vigilance task, reporting that novice learners produced greater mental activity than experienced learners and were thus assumed to have expended greater cognitive effort; this study also concluded that cerebral blood flow speed can be used to measure the amount of cerebral resources used to complete a task. Similarly, Durantin et al. (2014) looked at the piloting of remotely operated vehicles and compared fNIRS with two well-established measures of mental workload (heart-rate variability and subjective self-reports) and found that fNIRS proved effective at measuring mental overload detection.

Evidence supporting the possibility of more self-regulating technologies for managing cognitive load comes from studies aiming to optimise workload and learning in real time through measuring extended periods of boredom or cognitive overload. Participants in one such study conducted path planning in a high-frequency interaction environment for multiple unmanned aerial vehicles (“drones”) in a simulated environment. It was found that fNIRS was able to detect variations in task difficulty by changing the task difficulty

dynamically, usually by adding or removing the number of vehicles to be monitored simultaneously in response to measurement made with fNIRS (Afergan et al., 2014). This approach reduced the operator failure rate by 35% whilst creating conditions where they also exhibited higher levels of alertness for addressing problems, supporting findings from earlier similar studies using military flight simulators (e.g., Huttunen, Keränen, Väyrynen, Pääkkönen, & Lenio, 2011).

Although increasing cognitive load produces different brain network responses in younger than in older individuals, the key principle behind both fMRI and fNIRS is the same; that selective brain regions show progressively increased activity as cognitive load is increased (typically in the prefrontal cortex), but it is recognised that other brain regions are also involved depending on the nature of the task (O'Hare, Lu, Houston, Brookheimer, & Sowell, 2008) and that, generally, the longer and more intense the learning stimulus the longer the duration required for recovery in cortical haemodynamics between learning sessions (Leff et al., 2011). fNIRS therefore seems able to detect workload in a wide range of different contexts that are not limited to a specific type of working memory and has been used in studies where aircrew were piloting unmanned air vehicles (Ayaz, Cakir, et al., 2012; Ayaz, Shewokis, et al., 2012), when participants were part of a human-robot team (Solovey et al., 2011), when evaluating complex visual tasks (Peck, Yuksel, Ottley, Jacob, & Chang, 2013), and whilst driving vehicles (Tsunashima & Yanagisawa, 2009). One study of groups during a driving video-game found that the left and right prefrontal cortex differentiate between intrinsic and extrinsic cognitive load according to the predictability of events such as manoeuvring a vehicle (e.g., turning). Regardless of whether cognitive load was able to be attributed to intrinsic or extrinsic factors, this study also identified additional different brain regions involved in preparation for action, although load type did affect activation patterns (Liu, Saito, & Oi, 2012).

Earlier fNIRS studies have shown that if load increases to the point of reduced task performance, the signal changes in ways that suggest a capacity-limited response (e.g., Callicott et al., 1999). However, the interpretation of such measures requires care, and the relationship between task difficulty, mental effort, and brain activity is non-linear and subtle (Jaeggi et al., 2003). Additionally, cognitive load produces different brain network responses in younger than in older individuals (O'Hare et al., 2008), and so maturation effects – at present poorly understood – should also be accommodated in future measures.

Physiological measures may therefore claim high reliability but pose difficulties of construct validity because their outputs may be due to a range of factors that are exposed to problems of multi-causality. Despite the ability of fNIRS to track the functional pathways of the brain that mediate the maintaining and manipulating of attention, one problem with current research is that there is no standardised approach to extracting features from the signal or a prevailing consensus about which features of the fNIRS signal may result in the highest levels of accuracy (Peck, Afergan, Yuksel, Lalooses, & Jacob, 2014). Researchers also do not know in advance whether a period of interaction should be producing high or low mental workload.

Novice performance often falls at difficult levels whilst expert users show a strong increase, suggesting that fNIRS may have valuable potential for establishing user expertise in real

time as part of an adaptive system that could calibrate instructional content based on the sensed expertise of the learner (Bunce et al., 2011). So whilst fNIRS is able to measure signal changes over time with great accuracy, its use and the interpretation of its signal outputs requires care (as with EEG and FMRI), not least because the biological response of the brain to neurological changes is more sluggish, as changes in the oxygenation and deoxygenation of blood take time to reach brain tissues (Peck et al., 2014).

Individual differences

The influence of individual differences on learning task performance is important because it could significantly affect the interpretation of the role and impact of cognitive load in working memory. Evaluation that relies on verbal feedback can be influenced by individual preference and expectation, cultural biases in particular fields, or resistance to change (Anderson et al., 2011).

Working memory capacity is strongly influenced by individual differences in controlling attention. The capacity of working memory, or at least of its central executive element, is therefore subject to volitional control and is variable between individuals. However, although individuals possess different resources for mental processing, they may or may not choose to make use of them in any given situation, as most experienced educators know (Barrett, Tugade, & Engle, 2004). The cognitive load and participant confidence of e-learners as they completed an asynchronous task showed that procedural and subject knowledge or skill are not always enough to bring about successful learning and that an individual's self-efficacy and motivation are also important, possibly more so (Martin & Vallance, 2008; McQuaid, 2010).

Educators and researchers are familiar with the phenomenon whereby increases in cognitive load are not always accompanied by a decrease in task performance or an increase in time on task, as learning may be sustained by drawing on resources such as effort, commitment to task, and motivation (Brünken, Plass, & Moreno, 2010; Hockey, 1997; Peck et al., 2014). Performance on long tasks is known to decrease with time on task, but there has been disagreement about the substantive cause of this well-known effect, as decreases in vigilance have also been known to appear in tasks lasting less than 10 min. (Matthews, Davies, & Holley, 1993; Robertson, Manly, Andrade, Baddeley, & Yiend, 1997). Cognitive load theory appears to explain attentional intensity more successfully than arousal theory (see Stroh, 1971, for an overview), where changes in vigilance are assumed to be associated mainly with long-duration tasks. Several studies have argued that vigilance is determined more by resource demands than by task duration, and subjective alertness may therefore be a useful index of the availability of cognitive resources (Matthews et al., 1993; Robertson et al., 1997; Smit et al., 2004; Stroh, 1971). Similarly, although the combination of the effects of task difficulty such as pressure of time and alertness when executing a memory task have been shown to have an overall effect, an important variable tends to be individual differences, including alertness (Galy, Cariou, & Mélan, 2012).

It is therefore important to take due account of the large individual differences found in fMRI research, especially with regard to differences in processing at capacity limits. These differences are thought to most likely result from the more efficient use of resources that is

especially pronounced in high-performing individuals. It is unclear why these differences are present, but it is thought to hinge on high-performing individuals' better differentiation between relevant and irrelevant information (Jaeggi et al., 2007; Unsworth et al., 2004).

Other studies have concluded that individuals with higher IQs have more cognitive resources available for processing information and so experience lower cognitive demands than less intelligent individuals engaged on the same task, probably as a result of differences in the rates of information processing (Fink & Neubauer, 2005). High workload has also been found to be more detrimental to individuals with low cognitive abilities than to those with high cognitive abilities, in both traditional learning settings and contexts featuring blended or technology-enhanced learning. Individuals with low cognitive abilities also appear to be more sensitive to workload during such learning than those with high cognitive abilities, and there appears to be a general link between intelligence and learning in real-time dynamic decision making tasks (Fink & Neubauer, 2005; Gonzalez, 2005). A negative correlation has been found to exist between brain activity under cognitive load and intelligence, suggesting that individuals with high IQ scores process information more quickly and perhaps differently and may use more optimal mental strategies than individuals with low IQ scores, even when no reasoning tasks are involved, perhaps due to differences in the central executive in rates of controlling, switching and focussing attention and in inhibiting irrelevant processes (Fink & Neubauer, 2005; Gibbons & Stahl, 2010; Jaušovec & Jaušovec, 2004).

However, the concept of a central executive which controls this focus of attention in many different situations remains the least well understood element of the working-memory system (Baddeley & Hitch, 1994), and the outcome of some earlier studies even undermines the important related notion that the limitation of working memory relates to the number of "chunks" that can be held in memory (see Unsworth et al., 2004). It is possible that such findings mask the operation of more complex underlying cognitive structures and prior learned responses as, for example, when attempting to account for which cognitive mechanism is responsible for the longer reading times spent on words that convey more information. Frank (2013) proposes that this effect is due to the reducing of meaning uncertainty (reducing entropy) when reading words in sentences, but a mechanistic model has not to date been established to develop this.

Sex-related differences in cognition have also been found for some spatial measures, where a male advantage was present (Voyer, Voyer, & Bryden, 1995), whereas a female advantage has been found for some verbal measures (Crossley, D'Arcy, & Rawson, 1997; Kramer, Delis, & Daniel, 1988; Norman, Evans, Miller, & Heaton, 2000; Weiss et al., 2006), when using object location memory measures (see Sykes Tottenham, Saucier, Elias, & Gutwin, 2003; Voyer, Postma, Brake, & Imperato-McGinley, 2007), for spatial cognition, language, and memory (Lejbak, Crossley, & Vrbancic, 2011). These findings could have important implications for the use of different kinds of e-learning or blended-learning solutions, and it may be necessary to take these into account when instructors are designing or using multimedia or web-based resources. Other working memory tasks have also shown gender-related differences in neural activity, although findings have been inconsistent, and some studies have found contradictory outcomes, suggesting that females and males may have used different strategies when completing

tasks and that older women had more difficulty completing the task, leading the authors to speculate whether oestrogen levels played a part in this (Goldstein et al., 2005; Lejbak et al., 2011; Nagel, Ohannessian, & Cummins, 2007; Speck et al., 2000).

We might reasonably expect that learners with different levels of content expertise will probably report different levels of both cognitive load and task complexity for a given task because of differences in their automated germane prior learning and schema acquisition. The level of expertise does appear to influence brain response for some complex tasks, but as yet we have no reliable way of measuring the extent of an individual's schema resources and the degree to which such automated prior learning mediates cognitive load during an individual learning task (Ayaz, Cakir, et al., 2012; Brünken, Seufert, & Paas, 2010). The prospect of obtaining indirect objective measures may lead educators to be less concerned with the degree of individual schema acquisition, but it is likely that they would be interested in metrics that provide insights into the implications of different kinds of prior learning. Reasonably reliable and valid indirect measures may prove to have more immediate value to educators than direct measures with higher levels of complexity and uncertainty. However, this may be changing as fNIRS systems provide increasingly sensitive interactive systems for mediating user experiences and subjective alertness. It may also be the case that training students in more generic skills such as attention switching would be attractive, providing such skills were able to be transferred across different tasks and knowledge domains, although educators may need to remain alert to persistent underlying influences of gender and intelligence.

Conclusion

Educators are understandably interested in promoting progressively higher levels of success in learning outcomes whatever approach to educational pedagogy they adopt, and such levels, however these are defined, are generally the main focus of instruments assessing the effect of instruction. However, learning outcomes by themselves cannot be regarded as valid measures of cognitive load and, perhaps more importantly, do not necessarily offer sufficient guidance on which resources or approaches are best suited to maximising the future learning of an individual. As we have seen, the measurement of cognitive load needs to take account of factors such as motivation, self-concept, and engagement with the task that have not yet featured strongly in empirical or theoretical work in this area, and in the absence of overarching metrics in this area the professional judgement of educators seems likely to retain a significant role.

Cognitive load measurements are also relative and transient and subject to a number of individual and empirical factors that vary over time. Cognitive load therefore cannot be seen as a constant factor related only to objective features of instructional format or content, and more developmental or longitudinal studies might add significantly to our understanding as a result. Similarly, the number of separate interacting elements needed to solve a task is an index of task complexity (intrinsic cognitive load), and increasing task complexity is likely to be reflected in greater demands on cognitive resources, but objective measures of task complexity (e.g., by cognitive task analysis) are not currently used in cognitive load research. Intrinsic load for a given task is also affected by individual learner aptitudes such as prior expertise (Seufert, Jänen, & Brünken, 2007), and it may

therefore be more useful to define task complexity in terms of the amount of information that has to be extracted from a given information source in relation to the goal of the learning task, and better statistical measures such as time series analysis have been proposed to help with this in future research (Brünken, Seufert, & Paas, 2010).

What appears necessary for significant advances to be made in cognitive load research are more unobtrusive and reliable means of continually monitoring different types of cognitive demand in authentic learning situations over extended periods of time and especially in complex or loosely structured domains. Measuring mental load has become the single most problematic and important issue in cognitive load theory but:

To date, there is no model integrating the role of learner characteristics, such as individual differences in prior knowledge, working memory capacity, and domain-specific abilities that can help predict the relative intrinsic difficulty of the material for a specific learner in a specific situation. (Brünken, Plass, & Moreno, 2010, pp. 257)

Cognitive load theory argues that learning helps promote schema development and acquisition, but this too is as yet unproven. Direct empirical evidence for the cognitive mechanisms by which schema acquisition comes about is lacking. As yet, we also have no instruments to directly measure a learner's working memory resources. We are still working towards a full understanding of how cognitive load relates to specific forms of knowledge representation, although we have some evidence that these are closely linked and that the type of skills and knowledge that a learner acquires depend upon the particular kinds of mental representations promoted by instruction (Schnotz, Boeckheler, & Grzondziel, 1999; Wallen, Plass, & Brünken, 2005).

We also do not yet know what kinds of mental processes in instructional design will best promote schema acquisition or of what kind. For example, the introduction of an appropriate image into textual learning materials tends to promote greater understanding (Mayer, 2001), but cognitive load theory has difficulty in explaining whether such an arrangement causes a decrease in the overall cognitive load, increases the germane load, decreases the extraneous load, or enhances helpful schema formation. The disentangling of different kinds of load and successfully measuring them in valid and reliable ways remains a challenge for the research field. Understanding the basic foundations of working memory is also an important priority for cognitive load theory as this concept is central to its structure, but as yet the theory offers no explicit assumptions about its architecture or operation.

Some writers seriously doubt that it will be possible to find the holy grail of cognitive load theory, reliable and valid individual measures for the three types of cognitive load (see Kirschner et al., 2011), although others report some limited progress (Leppink, Paas, Van der Vleuten, Van Gog, & Van Merriënboer, 2013). In general, there has been limited success in disentangling and measuring different types of cognitive load, and current research in this area features considerable discrepancies in the wording of cognitive load measures, when they are collected during an intervention and how "efficiency" is defined and used. Securing valid and reliable metrics will need to take due account of the importance of the sensitivity of instruments used to measure cognitive load and also the importance of the diagnosticity of them – that is, the ability of an

instrument to discriminate between the different types of mental/cognitive load – and reduce the intrusiveness of the instrument, as an overly intrusive measure will interfere more with the primary task performance (Yuan, Steedle, Shavelson, Alonzo, & Opezzo, 2006; Wiebe, Roberts, & Behrend, 2010).

At the moment, many performance test results do not correlate well with the subjective measures used, and one or other of the two effects (load and performance) will appear, but not both, and they occasionally conflict each other or confound the theoretical argument being applied. Attempts to discriminate between and measure more than one type of load have been highly problematical and so far have tended to fail because outcomes have either been inconsistent or highly correlated. However, the theoretical and practical problems to be solved are clearly identified, and despite all the remaining challenges, there are indications that instruments such as fNIRS may offer ways towards empirical and objective, valid, and reliable individual measures of cognitive load, and the field remains a rich source of opportunity for the exploration and understanding of cognition.

References

- Afergan, D., Peck, E. M., Solovey, E. T., Jenkins, A., Hincks, S. W., Brown, E. T.,...Jacob, R. J. K. (2014). Dynamic difficulty using brain metrics of workload. In CHI'14 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 3797–3806). New York, NY: ACM Press.
- Anderson, E.W., Potter, K. C., Matzen, L. E., Shepherd, J. F., Preston, G. A., & Silva, C. T. (2011). A user study of visualisation effectiveness using EEG and cognitive load. *Computer Graphics Forum*, 30, 791–800.
- Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, MA: University Press.
- Antonenko, P. D., & Neiderhauser, D. S. (2010). The influence of leads on cognitive load and learning in a hypertext environment. *Computers in Human Behavior*, 26, 140–150.
- Argyris, C. (1976). Theories of action that inhibit individual learning. *American Psychologist*, 31, 636–654.
- Ayaz, H., Cakir, M. P., Izzetoglu, K., Curtin, A., Shewokis, P. A., Bunce, S., & Onaral, B. (2012, April). Monitoring expertise development during simulated UAV piloting tasks using optical brain imaging. Paper presented at the IEEE Aerospace conference, BigSky, MN.
- Ayaz, H., Shewokis, P. A., Bunce, S., Izzetoglu, K., Willems, B., & Onaral, B. (2012). Optical brain monitoring for operator training and mental workload assessment. *NeuroImage*, 59, 36–47.
- Ayres, P. (2006). Using subjective measures to detect variations of intrinsic cognitive load within problems. *Learning and Instruction*, 16, 389–400.
- Ayres, P., & Paas, F. (2012). Cognitive load theory: New directions and challenges. *Applied Cognitive Psychology*, 26, 827–832.
- Ayres, P., & Van Gog, T. (2009). State of the art research into cognitive load theory. *Computers in Human Behavior*, 25, 253–257.
- Baddeley, A. D. (1986). *Working memory*. Oxford, UK: Oxford University Press.
- Baddeley, A. D. (1992). *Working Memory*. *Science*, 255, 556–559.

- Baddeley, A. D., & Hitch, G. J. (1994). Developments in the concept of working memory. *Neuropsychology*, 8, 485–493.
- Baddeley, A. D., & Logie, R. H. (1999). Working memory: The multiple-component model. In A. Miyake & P. Shah (Eds.), *Models of working-memory: Mechanisms of active maintenance and executive control* (pp. 28–61). Cambridge, UK: Cambridge University Press.
- Barrett, L. F., Tugade, M. M., & Engle, R. W. (2004). Individual differences in working memory capacity and dual-process theories of the mind. *Psychological Bulletin*, 130, 553–573.
- Barrouillet, P., Bernardin, S., & Camos, V. (2004). Time constraints and resource sharing in adults' working memory spans. *Journal of Experimental Psychology: General*, 133, 83–100.
- Barrouillet, P., Bernardin, S., Portrat, S., Vergauwe, E., & Camos, V. (2007). Time and cognitive load in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33, 570–585.
- Blanshard, B. (1962). *Reason and analysis*. London, UK: George Allen & Unwin.
- Block, R. A., Hancock, P. A., & Zakay, D. (2010). How cognitive load affects duration judgments: A meta-analytic review. *Acta Psychologica*, 134, 330–343.
- Bratfisch, O., Borg, G., & Dornic, S. (1972). Perceived item-difficulty in three tests of intellectual performance capacity (Technical Report No. 29). Stockholm, Sweden: Institute of Applied Psychology.
- Brisson, J., Mainville, M., Mailloux, D., Bealieu, C., Serres, J., & Sirois, S. (2013). Pupil diameter measurement errors as a function of gaze direction in corneal eyetrackers. *Behavior Research Methods*, 45, 1322–1331.
- Brünken, R., Plass, J. L., & Leutner, D. (2003). Direct measurement of cognitive load in multimedia learning. *Educational Psychologist*, 38, 53–61.
- Brünken, R., Plass, J. L., & Moreno, R. (2010). Current issues and open questions in cognitive load research. In J. L. Plass, R. Moreno, & R. Brünken (Eds.), *Cognitive load theory* (pp. 253–272). Cambridge, NY: Cambridge University Press.
- Brünken, R., Seufert, T., & Paas, F. (2010). Measuring cognitive load. In J. L. Plass, R. Moreno, & R. Brünken (Eds.), *Cognitive load theory* (pp. 181–202). Cambridge, NY: Cambridge University Press.
- Brünken, R., Steinbacher, S., Plass, J. L., & Leutner, D. (2002). Assessment of cognitive load in multimedia learning using dual-task methodology. *Experimental Psychology*, 49, 109–119.
- Brünken, R., Steinbacher, S., Schnotz, W., & Leutner, D. (2001). Mentale Modelle und Effekte der Präsentations und Abrufkodalität beim Lernen mit Multimedia [Mental models and the effects of presentation and retrieval mode in multimedia learning]. *Zeitschrift für Pädagogische Psychologie*, 15, 15–27.
- Bunce, S. C., Izzetoglu, K., Ayaz, H., Shewokis, P., Issetoglu, M., Pourrezaei, K., & Onaral, B. (2011). Implementation of fNIRS for monitoring levels of expertise and mental workload. In D. D. Schmorrow & C. M. Fidopiastis (Eds.), *Foundations of augmented cognition. Directing the future of adaptive systems* (pp. 13–22). Berlin, Germany: Springer.
- Çakir, M. P., Ayaz, H., Izzetoglu, M., Shewokis, P. A., Izzetoglu, K., & Onaral, B. (2011). Bridging brain and educational sciences: An optical brain imaging study of visuospatial reasoning. *Procedia – Social and Behavioral Sciences*, 29, 300–309.
- Callan, M. J., Sutton, R. M., & Dovale, C. (2010). When deserving translates into causing: The

- effect of cognitive load on immanent justice reasoning. *Journal of Experimental Social Psychology*, 46, 1097–1100.
- Callicott, J. H., Mattav, V. S., Bertolino, A., Finn, K., Coppola, R., Frank, J. A.,...Weinberger, D. R. (1999). Physiological characteristics of capacity constraints in working memory as revealed by functional MRI. *Cerebral Cortex*, 9, 20–26.
- Carswell, C. M. (2005). Assessing mental workload during laparoscopic surgery. *Surgical Innovation*, 12, 80–90.
- Chance, B., Anday, E., Nioka, S., Zhou, S., Hong, L., Worden, K.,...Thomas, R. (1998). A novel method for fast imaging of brain function, non-invasively, with light. *Optics Express*, 2, 411–423.
- Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction*, 8, 293–332.
- Chandler, P., & Sweller, J. (1992). The split-attention effect as a factor in the design of instruction. *British Journal of Educational Psychology*, 62, 233–246.
- Chang, C., & Yang, F. (2010). Exploring the cognitive loads of high-school students as they learn concepts in web-based environments. *Computers & Education*, 55, 673–680.
- Chen, S., & Epps, J. (2013). Automatic classification of eye activity for cognitive load measurement with emotion interference. *Computer Methods and Programs in Biomedicine*, 110, 111–124.
- Chen, S., Epps, J., Ruiz, N., & Chen, F. (2011, February). Eye activity as a measure of human mental effort in HCI. Paper presented at the International Conference on Intelligent User Interfaces (IUI'11), Palo, Alto, CA.
- Cierniak, G., Scheiter, K., & Gerjets, P. (2009). Explaining the split-attention effect: Is the reduction of extraneous cognitive load accompanied by an increase in germane cognitive load? *Computers in Human Behavior*, 25, 315–324.
- Coe, R., Aloisi, C., Higgins, S., & Major, L. E. (2014). What makes great teaching? Review of the underpinning research. London, UK: Sutton Trust. Retrieved from <http://www.suttontrust.com/wp-content/uploads/2014/10/What-Makes-Great-Teaching-REPORT.pdf>
- Cowan, N. (2000). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, 24, 87–185.
- Coyle, S., Ward, T., Markham, C., & McDarby, G. (2004). On the suitability of near-infrared (NIR) systems for next-generation brain-computer interfaces. *Physiological Measurement*, 25, 815–822.
- Crossley, M., D'Arcy, C., & Rawson, N. S. (1997). Letter and category fluency in community-dwelling Canadian seniors: A comparison of normal participants to those with dementia of the Alzheimer or vascular type. *Journal of Clinical and Experimental Neuropsychology*, 19, 52–62.
- Davies, P. (2000). The relevance of systematic reviews to educational policy and practice. *Oxford Review of Education*, 26, 365–378.
- De Jong, T. (2010). Cognitive load theory, educational research, and instructional design: Some food for thought. *Instructional Science*, 38, 105–134.
- DeLeeuw, K. E., & Mayer, R. E. (2008). A comparison of three measures of cognitive load: Evidence for separable measures of intrinsic, extraneous, and germane load. *Journal of Educational Psychology*, 100, 223–234.
- Durant, G., Gagnon, J.-F., Temblay, S., & Dehaes, F. (2014). Using near infrared spectroscopy and heart rate variability to detect mental overload. *Behavioural Brain*

Research, 259, 16–23.

- Dutke, S., & Rinck, M. (2006). Multimedia learning: Working memory and the learning of word and picture diagrams. *Learning and Instruction, 16*, 526–537.
- Eilam, B., & Poyas, Y. (2008). Learning with multiple representations: Extending multimedia learning beyond the lab. *Learning and Instruction, 18*, 368–378.
- Ferrari, M., & Quaresima, V. (2012). A brief review on the history of human functional near-infrared spectroscopy (fNIRS) development and fields of application. *NeuroImage, 63*, 921–935.
- Fink, A., & Neubauer, A. C. (2005). Individual differences in time estimation related to cognitive ability, speed of information processing and working memory. *Intelligence, 33*, 5–26.
- Fishburn, F. A., Norr, M. E., Medvedev, A. V., & Vaidya, C. J. (2014). Sensitivity of fNIRS to cognitive state and load. *Frontiers of Human Neuroscience, 8*:76.
doi:10.3389/fnhum.2014.00076
- Frank, S. L. (2013). Uncertainty reduction as a measure of cognitive load in science comprehension. *Topics in Cognitive Science, 5*, 475–494.
- Freeman, W. J., Ahlfors, S. P., & Menon, V. (2009). Combining fMRI with EEG and MEG in order to relate patterns of brain activity to cognition. *International Journal of Psychophysiology, 73*, 43–52.
- Galy, E., Cariou, M., & Mélan, C. (2012). What is the relationship between mental workload factors and cognitive loads types. *International Journal of Psychophysiology, 83*, 269–275.
- Gavens, N., & Barrouillet, P. (2004). Delays of retention, processing efficiency, and attentional resources in working memory span development. *Journal of Memory and Language, 51*, 644–657.
- Gerjets, P., Scheiter, K., & Cierniak, G. (2009). The scientific value of cognitive load theory: A research agenda based on the structuralist view of theories. *Educational Psychology Review, 21*, 43–54.
- Gibbons, H., & Stahl, J. (2010). Cognitive load reduces visual identity negative priming by disabling the retrieval of task-inappropriate prime information: An ERP study. *Brain Research, 1330*, 101–113.
- Goldstein, J. M., Jerram, M., Poldrack, R., Anagnoson, R., Breiter, H. C., Makris, N.,...Seidman, L. J. (2005). Sex differences in prefrontal cortical brain activity during fMRI of auditory verbal working memory. *Neuropsychology, 19*, 509–519.
- Gonzalez, C. (2005). Task workload and cognitive abilities in dynamic decision making. *Human Factors, 47*, 92–101.
- Gough, D., Oliver, S., & Thomas, J. (2013). Learning from research: Systematic reviews for informing policy decisions: A quick guide (A paper for the Alliance for Useful Evidence). London, UK: Nesta. Retrieved from
<http://www.alliance4usefulevidence.org/assets/Alliance-final-report-08141.pdf>
- Grandchamp, R., Braboszcz, C., & Delorme, A. (2014). Oculometric variations during mind wandering. *Frontiers in Psychology, 5*. doi:10.3389/fpsyg.2014.00031
- Greef, T., Lafeber, H., Oostendorp, H., & Lindenberg, J. (2009). Eye movement as indicators of mental workload to trigger adaptive automation. In D. D. Schmorrow, I. V. Estabrooke, & M. Grootjen (Eds.), *Foundations of augmented cognition. Neuroergonomics and operational neuroscience* (pp. 219–228). Berlin, Germany: Springer.

- Grimes, D., Tan, D. S., Hudson, S. E., Shenoy, P., & Rao, R. P. N. (2008). Feasibility and pragmatics of classifying working memory load with and Electroencephalograph. Retrieved from <http://research.microsoft.com/pubs/64270/chi2008-cogload.pdf>
- Halford, G. S., Baker, R., McCredden, J. E., & Bain, J. D. (2005). How many variables can humans process? *Psychological Science*, 16, 70–76.
- Higgins, J. P. T., & Green, S. (2011). *Cochrane handbook for systematic reviews of interventions* (Version 5.1.0 [updated March 2011]). Retrieved from www.cochrane-handbook.org
- Hockey, G. R. J. (1997). Compensatory control in the regulation of human performance under stress and high workload: A cognitive-energetical framework. *Biological Psychology*, 45, 73–93.
- Howard-Jones, P. (2014). *Neuroscience and education: A review of educational interventions and approaches informed by neuroscience*. Bristol, UK: Education Endowment Foundation.
- Huttunen, K., Keränen, H., Väyrynen, E., Pääkkönen, R., & Lenio, T. (2011). Effect of cognitive load on speech prosody in aviation: Evidence from military simulator flights. *Applied Ergonomics*, 42, 348–357.
- Irwin, D. E., & Thomas, L. E. (2010). Eyeblinks and cognition. In V. Coltheart (Ed.), *Tutorials in visual cognition* (pp. 121–141). New York, NY: Psychology Press.
- Jacob, R. J. K., & Karn, K. S. (2003). Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. In J. Hyona, R. Radasch, & H. Deubels (Eds.), *The mind's eye: Cognitive and applied aspects of eye movement research* (pp. 573–603). Oxford, UK: Elsevier Science.
- Jaeggi, S. M., Buschkuhl, M., Etienne, A., Ozdoba, C., Perrig, W. J., & Nirkko, A. C. (2007). On how high performers keep cool brains in situations of cognitive overload. *Cognitive, Affective, & Behavioral Neuroscience*, 7, 75–89.
- Jaeggi, S. M., Seewer, R., Nirkko, A. C., Eckstein, D., Schroth, G., Groner, R., & Gutbrod, K. (2003). Does excessive memory load attenuate activation in the prefrontal cortex? Load-dependant processing in single and dual tasks: Functional magnetic resonance imaging study. *NeuroImage*, 19, 210–225.
- Jaušovec, N., & Jaušovec, K. (2004). Differences in induced brain activity during the performance of learning and working-memory tasks related to intelligence. *Brain and Cognition*, 54, 65–74.
- Johnstone, A. H. (1997). Chemistry teaching – Science or alchemy? *Journal of Chemical Education*, 74, 262–268.
- Kalyuga, S. (2009). *Managing cognitive load in adaptive multimedia learning*. London, UK: Information Science Reference.
- Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist*, 38, 23–31.
- Kalyuga, S., Chandler, P., & Sweller, J. (1998). Levels of expertise and instructional design. *Human Factors*, 40, 1–17.
- Kalyuga, S., & Sweller, J. (2005). Rapid dynamic assessment of expertise to improve the efficiency of adaptive e-learning. *Educational Technology Research and Development*, 53, 83–93.
- Khawaja, M. A., Ruiz, N., & Chen, F. (2007, November). Potential speech features for cognitive load measurement. Paper presented at the OZCHI conference, Adelaide, Australia.

- Kirschner, P. A., Ayres, P., & Chandler, P. (2011). Contemporary cognitive load theory research: The good, the bad and the ugly. *Computers in Human Behaviour*, 27, 99–105.
- Klingner, J., Kumar, R., & Hanrahan, P. (2008). Measuring the task-evoked pupillary response with a remote eye tracker. In *ETRA '08: Proceedings of the 2008 Symposium on Eye Tracking Research & Applications* (pp. 69–72). New York, NY: ACM.
doi:[10.1145/1344471.1344489](https://doi.org/10.1145/1344471.1344489)
- Klingner, J., Tversky, B., & Hanrahan, P. (2010). Effects of visual and verbal presentation on cognitive load in vigilance, memory, and arithmetic tasks. *Psychophysiology*, 48, 323–332.
- Kramer, A. F. (1991). Physiological metrics of mental workload: A review of recent progress. In D. L. Damos (Ed.), *Multiple-task performance* (pp. 279–328). London, UK: Taylor & Francis.
- Kramer, J. H., Delis, D. C., & Daniel, M. (1988). Sex differences in verbal learning. *Journal of Clinical Psychology*, 44, 907–915.
- Leff, D. R., Orihuela-Espina, F., Elwell, C. E., Athanasiou, T., Delpy, D. T., Darzi, A. W., & Yang, G.-Z. (2011). Assessment of the cerebral cortex during motor task behaviours in adults: A systematic review of functional near infrared spectroscopy (fNIRS) studies. *NeuroImage*, 54, 2922–2936.
- Lejbak, L., Crossley, M., & Vrbancic, M. (2011). A male advantage for spatial and object but not verbal working memory using the n-back task. *Brain and Cognition*, 76, 191–196.
- Leppink, J., Paas, F., Van der Vleuten, C. P. M., Van Gog, T., & Van Merriënboer, J. J. G. (2013). Development of an instrument for measuring different types of cognitive load. *Behavior Research Methods*, 45, 1058–1072.
- Lin, T., Li, X., Wu, Z., & Tang, N. (2013). Automatic cognitive load classification using high-frequency interaction events: An exploratory study. *International Journal of Technology and Human Interaction*, 9(3), 73–88.
- Liu, T., Saito, H., Oi, M. (2012). Distinctive activation patterns under intrinsically versus extrinsically driven cognitive loads in prefrontal cortex: A near-infrared spectroscopy study using a driving video game. *Neuroscience Letters*, 506, 220–224.
- MacLeod, C. M. (1992). The Stroop task: The “Gold Standard” of attentional measures. *Journal of Experimental Psychology: General*, 121, 12–14.
- Martin, S. (2010). Teachers using learning styles: Torn between research and accountability? *Teaching and Teacher Education*, 26, 1583–1591.
- Martin, S. (2012). Does instructional format really matter? Cognitive load theory, multimedia and teaching English Literature. *Educational Research and Evaluation*, 18, 125–152.
- Martin, S., & Vallance, M. (2008). The impact of synchronous inter-networked teacher training in information and communication technology integration. *Computers & Education*, 51, 34–53.
- Massa, L. J., & Mayer, R. E. (2006). Testing the ATI hypothesis: Should multimedia instruction accommodate verbalizer-visualizer cognitive style? *Learning and Individual Differences*, 16, 321–335.
- Matthews, G., Davies, D. R., & Holley, P. J. (1993). Cognitive predictors of vigilance. *Human Factors*, 35, 3–24.
- Mayer, R. E. (2001). *Multimedia learning*. New York, NY: Cambridge University Press.
- Mayer, R. E. (2003). The promise of multimedia learning: Using the same instructional design methods across different media. *Learning and Instruction*, 13, 125–139.

- Mayer, R. E. (2008). Applying the science of learning: Evidence-based principles for the design of multimedia instruction. *American Psychologist*, 63, 760–769.
- Mayer, R. E. (2009). *Multimedia learning* (2nd ed.). New York, NY: Cambridge University Press.
- Mayer, R. E., & Moreno, R. (2002). Aids to computer-based multimedia learning. *Learning and Instruction*, 12, 107–119.
- McQuaid, J. W. (2010). Using cognitive load to evaluate participation and design of an asynchronous course. *The American Journal of Distance Education*, 24, 177–194.
- Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity to process information. *Psychological Review*, 63, 81–97.
- Miller, L. M., Chang, C. I., Wang, S., Beier, M. E., & Klisch, Y. (2011). Learning and motivational impacts of a multimedia science game. *Computers & Education*, 57, 1425–1433.
- Moray, N. (1979). *Mental workload: Its theory and measurement*. London, UK: Springer.
- Moreno, R. (2006). When worked examples don't work: Is cognitive load theory at an impasse? *Learning and Instruction*, 16, 170–181.
- Moreno, R., & Park, B. (2010). Cognitive load theory: Historical development and relation to other theories. In J. L. Plass, R. Moreno, & R. Brünken (Eds.), *Cognitive load theory* (pp. 9–28). New York, NY: Cambridge University Press.
- Murai, K., Hayashi, Y., Okazaki, T., Stone, L. C., & Mitomo, N. (2008). Evaluation of ship navigator's mental workload using nasal temperature and heart rate variability. In *Proceedings of IEEE International Conference on Systems, Man and Cybernetics* (pp. 228–232). Red Hook, NY: Curran Associates.
- Nagel, B. J., Ohannessian, A., & Cummins, K. (2007). Performance dissociation during verbal and spatial working memory tasks. *Perceptual and Motor Skills*, 105, 243–250.
- Niegemann, H. M. (2001). *Neue Lernmedien: Konzipieren, entwickeln, einsetzen* [New instructional media: Conceptualize, develop, implement]. Bern, Switzerland: Huber.
- Norman, M. A., Evans, J. D., Miller, W. S., & Heaton, R. K. (2000). Demographically corrected norms for the California Verbal Learning Test. *Journal of Clinical and Experimental Neuropsychology*, 22, 80–94.
- O'Brien, S. (2006). Eye-tracking and translation memory matches. *Perspectives: Studies in Translatology*, 14, 185–205.
- O'Hare, E. D., Lu, L. H., Houston, S. M., Brookheimer, S. Y., & Sowell, E. R. (2008). Neurodevelopmental changes in verbal working memory load-dependency: An fMRI investigation. *NeuroImage*, 42, 1678–1685.
- Paas, F. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology*, 84, 429–434.
- Paas, F., Renkel, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist*, 38, 1–3.
- Paas, F., Tuovinen, J., Tabbers, H., & Van Gerven, P. W. M. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist*, 38, 63–71.
- Paas, F., & Van Merriënboer, J. J. G. (1993). The efficiency of instructional conditions: An approach to combine mental effort and performance measures. *Human Factors*, 35, 737–743.
- Paas, F., & Van Merriënboer, J. J. G. (1994). Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive load approach. *Journal of Educational*

Psychology, 86, 122–133.

- Paivio, A. (1986). *Mental representations: A dual coding approach*. New York, NY: Oxford University Press.
- Palinko, O., Kun, A. L., Shyrovkov, A., & Heeman, P. (2010). Estimating cognitive load using remote eye tracking in a driving simulator. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications* (pp. 141–144). Austin, TX: ACM.
- Peck, E. M., Afergan, D., Yuksel, B. F., Lalooses, F., & Jacob, R. J. K. (2014). Using fNIRS to measure mental workload in the real world. In S. H. Fairclough & K. Gilleade (Eds.), *Advances in physiological computing (Human-Computer Interaction Series)* (pp. 117–140). London, UK: Springer.
- Peck, E. M., Yuksel, B. F., Ottley, A., Jacob, R. J. K., & Chang, R. (2013). Using fNIRS brain sensing to evaluate information visualisation interfaces. In *CHI 2013 proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 473–482). Vancouver, BC: Association for Computing Machinery.
- Plass, J. L., Chun, D. M., Mayer, R. E., & Leutner, D. (2003). Cognitive load in reading a foreign language text with multimedia aids and the influence of verbal and spatial abilities. *Computers in Human Behavior*, 19, 221–243.
- Pollock, E., Chandler, P., & Sweller, J. (2002). Assimilating complex information. *Learning and Instruction*, 12, 61–86.
- Popper, K. (1959). *The logic of scientific discovery*. New York, NY: Basic Books.
- Popper, K. (1963). *Conjectures and refutations*. London, UK: Routledge.
- Portrat, S., Camos, V., & Barrouillet, P. (2009). Working memory in children: A time-constrained functioning similar to adults. *Journal of Experimental Child Psychology*, 102, 368–374.
- Reid, N. (2008). A scientific approach to the teaching of chemistry. What do we know about how students learn in the sciences, and how can we make our teaching match this to maximise performance? *Chemistry Education Research and Practice*, 9, 51–59.
- Repovš, G., & Baddeley, A. (2006). The multi-component model of working memory: Exploration in experimental cognitive psychology. *Neuroscience*, 139, 5–21.
- Robertson, I. H., Manly, T., Andrade, J., Baddeley, B. T., & Yiend, J. (1997). “Oops!” Performance correlates of everyday attentional failures in traumatic brain injured and normal subjects. *Neuropsychologia*, 35, 747–758.
- Ross, V., Jongen, E. M. M., Wang, W., Brijs, T., Brijs, K., Ruiters, R. A. C., & Wets, G. (2014). Investigating the influence of working memory capacity when driving behaviour is combined with cognitive load: An LCT study of young novice drivers. *Accident Analysis and Prevention*, 62, 377–387.
- Salomon, G. (1984). Television is “easy” and print is “tough”: The differential investment of mental effort in learning as a function of perceptions and attributions. *Journal of Educational Psychology*, 76, 647–658.
- Scharfenberg, F., & Bogner, F. X. (2013). Teaching gene technology in an outreach lab: Students’ assigned cognitive load clusters and the clusters’ relationship to learner characteristics, laboratory variables, and cognitive achievement. *Research in Science Education*, 43, 141–161.
- Schnotz, W. (2010). Reanalyzing the expertise reversal effect. *Instructional Science*, 38, 315–323.
- Schnotz, W., Boeckheler, J., & Grzondziel, H. (1999). Individual and co-operative learning with interactive animated pictures. *European Journal of Psychology of Education*, 14, 245–

- Schnotz, W., & Kürschner, C. (2007). A reconsideration of cognitive load theory. *Educational Psychology Review*, 19, 469–508.
- Schoor, C., Bannert, M., & Brünken, R. (2012). Role of dual task design when measuring cognitive load during multimedia learning. *Educational Technology Research and Development*, 60, 753–768.
- Seufert, T., Jänen, I., & Brünken, R. (2007). The impact of intrinsic cognitive load on the effectiveness of graphical help for coherence formation. *Computers in Human Behavior*, 23, 1055–1071.
- Shaw, T. H., Satterfield, K., Ramirez, R., & Finomore, V. (2013). Using cerebral haemovelocity to measure workload during spatialised auditory vigilance task in novice and experienced drivers. *Ergonomics*, 56, 1251–1263.
- Shi, Y., Choi, E. H. C., Ruiz, N. F., & Taib, R. (2007). Galvanic skin response (GSR) as an index of cognitive load. *ACM Digital Library. Extended abstracts on Human Factors in Computing Systems*, 2651–2656. Retrieved from dl.acm.org/citation.cfm?id=1241057
- Shuler, C. (2009). Pockets of potential: Using mobile technologies to promote children's learning. Retrieved from http://www.joanganzcooneycenter.org/wpcontent/uploads/2010/03/pockets_of_potential_1_.pdf
- Siegle, G. J., Steinhauer, S. R., & Thase, M. E. (2004). Pupillary assessment and computational modeling of the Stroop task in depression. *International Journal of Psychophysiology*, 52, 63–76.
- Smit, A. S., Eling, P. A. T. M., & Coenen, A. M. L. (2004). Mental effort causes vigilance decrease due to resource depletion. *Acta Psychologica*, 115, 35–42.
- Solovey, E. T., Lalooses, F., Chauncey, K., Weaver, D., Scheutz, M., Sassaroli, A.,...Jacob, R. J. K (2011). Sensing cognitive multitasking for a brain-based adaptive user interface. In *CHI 2011 proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 383–392). Vancouver, BC: Association for Computing Machinery.
- Speck, O., Ernst, T., Braun, J., Koch, C., Miller, E., & Chang, L. (2000). Gender differences in the functional organisation of the brain for working memory. *NeuroReport*, 11, 2581–2585.
- Strangman, G., Culver, J. P., Thompson, J. H., & Boas, D. A. (2002). A quantitative comparison of simultaneous BOLD fMRI and NIRS recordings during functional brain activation. *NeuroImage*, 17, 719–731.
- Stroh, C. M. (1971). *Vigilance: The problem of sustained attention*. Oxford, UK: Pergamon Press.
- Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12, 257–285.
- Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, 4, 295–312.
- Sweller, J. (1999). *Instructional design in technical areas*. Camberwell, Australia: ACER Press.
- Sweller, J., & Chandler, P. (1994). Why some material is difficult to learn. *Cognition and Instruction*, 12, 185–233.
- Sweller, J., Van Merriënboer, J. J. G., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, 10, 251–296.
- Sykes Tottenham, L., Saucier, D., Elias, L., & Gutwin, C. (2003). Female advantage for spatial location memory in both static and dynamic environments. *Brain and Cognition*, 53,

381–383.

- Tabbers, H. K., Martens, R. L., & Van Merriënboer J. J. G. (2000, February). Multimedia instructions and cognitive load theory: Split-attention and modality effects. Paper presented at the National Convention of the Association for Educational Communications and Technology, Long Beach, CA.
- Theeuwes, J., & Belopolsky, A. (2010). Top-down and bottom-up control of visual selection controversies and debate. In V. Coltheart (Ed.), *Tutorials in visual cognition* (pp. 67–92). New York, NY: Psychology Press.
- Towse, J. N., & Hitch, G. J. (1995). Is there a relationship between task demand and storage space in tests of working memory capacity? *The Quarterly Journal of Experimental Psychology*, 48, 108–124.
- Tsunashima, H., & Yanagisawa, K. (2009). Measurement of brain function of car driving using functional near-infrared spectroscopy (fNIRS). *Computational Intelligence and Neuroscience*. Retrieved from <http://dx.doi.org/10.1155/2009/164958>
- Unsworth, N., Schrock, J. C., & Engle, R. W. (2004). Working memory capacity and the Antisaccade Task: Individual differences in voluntary saccade control. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30, 1302–1321.
- Van Gerven, P.W. M., Paas, F., Van Merriënboer, J. J. G., & Schmidt, H. G. (2004). Memory load and the cognitive pupillary response in aging. *Psychophysiology*, 41, 167–174.
- Van Gog, T., Kester, L., Nievelstein, F., & Paas, F. (2009). Uncovering cognitive processes: Different techniques that can contribute to cognitive load research and instruction. *Computers in Human Behavior*, 25, 325–331.
- Van Gog, T., & Paas, F. (2008). Instructional efficiency: Revising the original construct in educational research. *Educational Psychologist*, 43, 16–26.
- Van Merriënboer, J. J. G., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology Review*, 17, 147–177.
- Van Orden, K. F., Limbert, W., Makeig, S., & Jung, T. (2009). Activity correlates of workload during a visuospatial memory task. *Human Factors*, 43, 111–121.
- Veenman, M. V. J., Prins, F. J., & Verheij, J. (2003). Learning styles: Self-reports versus thinking aloud measures. *British Journal of Educational Psychology*, 73, 357–372.
- Voyer, D., Postma, A., Brake, B., & Imperato-McGinley, J. (2007). Gender differences in object location memory: A meta-analysis. *Psychonomic Bulletin and Review*, 14, 23–38.
- Voyer, D., Voyer, S., & Bryden, M. P. (1995). Magnitude of sex differences in spatial abilities: A metaanalysis and consideration of critical variables. *Psychology Bulletin*, 117, 250–270.
- Wallen, E., Plass, J. L., & Brünken, R. (2005). The function of annotations in the comprehension of scientific texts: Cognitive load effects and the impact of verbal ability. *Educational Technology Research and Development*, 53, 59–71.
- Wauters, K., Desmet, P., & Van Den Noortgate, W. (2012). Item difficulty estimation: An auspicious collaboration between data and judgment. *Computers & Education*, 58, 1183–1193.
- Weiss, E. M., Ragland, J. D., Brensinger, C. M., Bilker, W. B., Deisenhammer, E. A., & Delazer, M. (2006). Sex differences in clustering and switching verbal fluency tasks. *Journal of the International Neuropsychological Society*, 12(4), 502–509.
- Whelan, R. R. (2007). Neuroimaging of cognitive load in instructional multimedia.

Educational Research Review, 2, 1–12.

- Wiebe, E. N., Roberts, E., & Behrend, T. S. (2010). An examination of two mental workload measurement approaches to understanding multimedia learning. *Computers in Human Behavior*, 26, 474–481.
- Wittrock, M. C. (1974). Learning as a generative process. *Educational Psychologist*, 11, 87–95.
- Yap, T. F., Ambikairajah, E., Choi, E., & Chen, F. (2009, April). Phase based features for cognitive load measurement system. Paper presented at the IEEE International Conference on Acoustics, Speech, and Signal Processing, Taipei, Taiwan.
- Yap, T. F., Epps, J., Ambikairajah, E., & Choi, E. H. C. (2011). Formant frequencies under cognitive load: Effects and classification. *EURASIP Journal on Advances in Signal Processing*, 2011:219253. doi:[10.1155/2011/219253](https://doi.org/10.1155/2011/219253)
- Yin, B., Chen, F., Ruiz, N., & Ambikairajah, E. (2008, March-April). Speech-based cognitive load monitoring system. Paper presented at the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, Las Vegas, NV.
- Young, M. S., & Stanton, N. A. (2002). It's all relative: Defining mental workload in the light of Annett's paper. *Ergonomics*, 45, 1018–1020.
- Yuan, K., Steedle, J., Shavelson, R., Alonzo, A., & Oppezzo, M. (2006). Working memory, fluid intelligence, and science learning. *Educational Research Review*, 1, 83–98.
- Zheng, R., & Cook, A. (2012). Solving complex problems: A convergent approach to cognitive load measurement. *British Journal of Educational Technology*, 43, 233–246.