



A Testing Load: Investigating Test Mode Effects on Test Score, Cognitive Load and Scratch Paper Use with Secondary School Students

James Pengelley¹ · Peter R. Whipp¹ · Nina Rovis-Hermann¹

Accepted: 23 May 2023 / Published online: 20 June 2023
© The Author(s) 2023

Abstract

The aim of the present study is to reconcile previous findings (a) that testing mode has no effect on test outcomes or cognitive load (Comput Hum Behav 77:1–10, 2017) and (b) that younger learners' working memory processes are more sensitive to computer-based test formats (J Psychoeduc Assess 37(3):382–394, 2019). We addressed key methodological limitations in past cognitive load research by employing a repeated measures design with 263, year 9 (aged 13–14) science students in Western Australia. Question difficulty (intrinsic cognitive load) and test mode (extraneous cognitive load) were manipulated to measure changes in test performance, cognitive load and scratch paper use on equivalent paper and computer-based versions of an Ohm's Law revision quiz. Hierarchical linear modelling indicated significantly higher paper-based test performance on difficult questions in addition to greater cognitive load and scratch paper use for all paper questions. Testing mode effects on test score, as well as both measures of cognitive load, were not significant when controlling for working memory capacity, although the testing mode*question difficulty interaction remained significant. Together, these results contradict previous findings that computer-based testing can be implemented without consequence for all learners. With the increased use of computer-based testing in national and international-level assessments, these findings warrant further research into the effect of different testing modes on school-aged students.

Keywords Testing mode · Cognitive load · Computer-based testing · Paper-based testing · Working memory

✉ James Pengelley
japengelley@gmail.com

¹ School of Education, College of Science, Health, Engineering and Education, Murdoch University, Murdoch, WA 6150, Australia

Introduction

The use of computers in Australian schools continues to increase (Hatzigianni et al., 2016; Vassallo & Warren, 2017), finding a place not only in learning environments but also in summative and high-stakes assessment. For instance, the National Assessment Program in Literacy and Numeracy (NAPLAN) in Australia is intended to be conducted entirely online by 2022 (School Curriculum and Standards Authority, 2014) whilst the Programme for International Student Assessment (PISA) has been transitioning to digital formats for nearly 16 years (OECD, 2010). The purported benefits of computer-based learning and assessment range from increased efficiency and accuracy for teachers in evaluating student performance (Prisacari & Danielson, 2017; Selwyn, 2014, 2016) to improved learning outcomes and test performance for students (Ackerman & Lauterman, 2012; Prisacari & Danielson, 2017).

Most research in this area has focused on whether the mode in which a test is taken (i.e., on computer or paper) leads to differences in testing outcomes (i.e., test scores, which validate primarily a student's retention of knowledge). However, testing mode (i.e., whether a student takes an assessment on computer or on paper) may affect school-aged students in inequitable ways based on their age and working memory capacity (Batka & Peterson, 2005; Bennett et al., 2008; Carpenter & Alloway, 2019). For instance, some studies point to subtle changes in cognitive load, metacognitive activity, and test-taking behaviours such as scratch paper. Prisacari and Danielson (2017) propose that when tackling some assessment questions, students may require additional support for their working out by adding key pieces of information they have memorised, sketching relationships between variables, or algebraic rearranging of formula. This is especially true for questions that include calculations (Bennett et al., 2008; Prisacari & Danielson, 2017). The reliance on spare paper to add additional steps required to arrive at a solution to a question is termed scratch-paper use.

Previous research highlights that experiences of an assessment question taken on computer compared to one taken on paper may not necessarily be reflected in differences in test scores (Galy et al., 2012; Mueller & Oppenheimer, 2014; Paas & Van Merriënboer, 1993; Prisacari & Danielson, 2017; Sidi et al., 2017). These observations warrant further investigation of computer-based teaching and assessment with school-aged students. In particular, there is a need for consideration of whether traditional measures of performance (i.e., test scores) are truly reflective of the subtle ways in which testing mode may provide a selective advantage to certain students (Martin, 2014).

To advance the scholarship in this area, this research compared the differences in (1) test scores, (2) subjective reports of cognitive load, and (3) test behaviour, in equivalent computer and paper-based versions of a grade 9 physics quiz. This research addressed the limitations evident in Prisacari and Danielson's (2017) study in that key sources of error (between-subjects measurements taken across multiple days) have been removed. Before outlining the study, a brief literature review is presented.

Literature Review

Cognitive Load in Learning Tasks

According to CLT, when the demands of a task exceed a learner's working memory capacity (WMC), the learner experiences an increase in cognitive load (CL), resulting in poorer test performance, poorer efficiency in answering questions, or poorer retention of content knowledge (Paas & Van Merriënboer, 1993; Sweller et al., 2019; Sweller, 1988, 2020). It should be acknowledged that CLT literature draws its theoretical base from a wide range of influences including theoretical models of cognitive architecture (Anderson, 1983), dual-channel models of working memory (Baddeley & Hitch, 1994) and Mayer's theory of multimedia learning (Mayer, 2003). However, this article refers primarily to the terminology of the tripartite model of cognitive load (Sweller et al., 1998).

This model proposes three types of cognitive load associated with a learning task. First, *intrinsic load* (IL) is associated with the complexity of the material being learned. Material that is high in *element interactivity* (EI) is said to be high in IL because the number of units of information intrinsic to the goals of the learning or assessment task that must be concurrently processed in working memory is relatively high (Likourezos et al., 2019). Recent developments to the conceptualization of IL account for the fact that the subjective experience of IL depends on individual learner characteristics (Kalyuga & Singh, 2016; Klepsch et al., 2017; Skulmowski & Xu, 2021). For instance, learners with high prior knowledge and high WMC are likely to report lower levels of IL on questions of equivalent difficulty (Beckmann, 2010; Klepsch et al., 2017; Naismith et al., 2015; Park et al., 2015). Furthermore, IL also operates as a function of a learner's motivation to engage with a task, or achieve a given instructional goal (Kalyuga & Singh, 2016; Schnotz & Kürschner, 2007). The implication of this is that learners who feel demotivated or who have low self-efficacy are less likely to engage in a learning or assessment task (Pajares, 2005) and thus report lower cognitive load.

The second type of load, *extraneous load* (EL), is associated with factors that do not relate to learning or content. Typically, this equates to load arising from instructional design or environment and is responsible for the majority of cognitive load effects (Sweller, 2020). In multimedia learning and assessment environments, EL is commonly caused by irrelevant information or digital distractions, often termed *seductive details* (Korbach et al., 2018), or causing a split-attention effect that arises when learners are required to split their attention between different sources of information (Sweller et al., 2019). EL and IL are additive to determine the total cognitive load experienced by a learner (De Jong, 2010; Jiang & Kalyuga, 2020; Kalyuga, 2011; Sweller, 2010; Sweller et al., 2011). In other words, when the demands of the content being learned (IL), coupled with demands of the instruction and learning environment (EL), exceed a learner's working memory capacity, students typically experience poorer test performance, poorer comprehension, and poorer attention control (Ayres, 2001, 2006; Debue & Van De Leemput, 2014; Mayes et al., 2001; Schmeck et al., 2015). The intrinsic

load of a given task is constant (for learners of equivalent expertise) whilst extraneous load on the other hand can be manipulated by design of learning and assessment tasks (Likourezos et al., 2019). Therefore, the goal of CLT is to identify ways to reduce EL, thus freeing cognitive resource that can be devoted to learning (see Fig. 1).

As Fig. 1 illustrates, the same learner (represented here by identical “bandwidth” provided by working memory resources in each scenario) is presented with an identical task (represented by the same amount of IL) under different conditions. In the first example (Fig. 1A), the learner’s working memory resources are overloaded and there are insufficient resources available to process, manipulate, or retain information. In the second example (Fig. 1B), working memory capacity is not exceeded. The remaining working memory resources can be devoted to successful task completion.

The third type of load proposed by the original CLT model, *germane load*, is associated with the cognitive resources devoted to dealing with the content of a task. This has since been incorporated into IL and has been removed from the tripartite model as a discrete type of load (Jiang & Kalyuga, 2020; Skulmowski & Xu, 2021), as it is no longer theorised to contribute to the overall load of a task (Sweller et al., 2019).

Cognitive Load in Assessment Tasks

Traditionally, CLT seeks to minimise total CL (by minimising EL and optimising IL (Sweller, 2018; Van Merriënboer et al., 2006)) and maximise residual working memory resources through selective use of instructional sequences. In doing so, incoming information can be processed with minimal error, and new schema can be formed in long-term memory (LTM) (Sweller, 2020). In applying CLT to assessment, it is reasonable to assume that, just as the success of a learning sequence is determined by the cognitive conditions under which information is processed as it enters LTM, so too should the success of assessment activities be determined by the cognitive conditions under which information is retrieved from LTM. Following descriptions of human cognitive architecture, students who are completing an assessment task must access existing schematic knowledge held in LTM, which must be manipulated in working memory to fit the context of a given assessment question (Sweller et al., 1998). Therefore, one might conclude that if the conditions

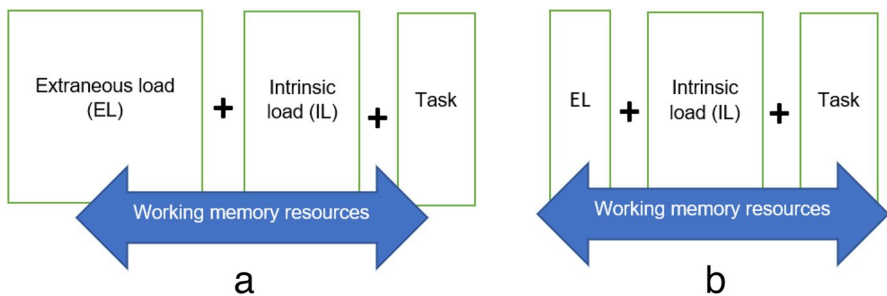


Fig. 1 The effect of increased extraneous load on the same learner completing the same task

under which students are assessed place learners under sufficient extraneous load, there may not be sufficient WMC to meet the demands of the assessment task.

Comparisons of Computer vs Paper-Based Mediums

With respect to the ways in which testing modality influences a student's experience in processing information, recent studies highlight that digitally presented information does not necessarily impede comprehension when compared to information presented on paper (Latini et al., 2020; Ronconi et al., 2022). However, digitally presented information has been shown to lead to longer reading time in school-aged male students (who also over-estimate their performance more than school-aged female students when reading on a computer) (Ronconi et al., 2022). Furthermore, on-screen information has also been shown to induce shallower processing than information presented on paper (Delgado & Salmerón, 2021; Delgado et al., 2018). Research by Latini et al. (2020) used eye-tracking measures to show that undergraduate students show more efficient eye gaze transitions between corresponding elements of textual and pictorial information when reading information presented on paper. This finding is thought to reflect the fact that students process on-screen information more shallowly (Annisette & Lafreniere, 2017) and leads them to over-estimate the extent to which they have understood key elements of a text (Ackerman & Lauterman, 2012; Sidi et al., 2017). In summary, whilst on-screen information may not necessarily always lead to differences in comprehension, there are subtle changes in student behaviour (such as completion time, eye tracking patterns, and self-evaluation of performance) that reflects differences in student experiences when encountering information on a computer.

In relation to cognitive load during computer or paper-based assessments, there is a lack of consensus around which mode is superior, with a number of studies suggesting that students achieve higher test scores on computer tasks (Ackerman & Lauterman, 2012; Noyes et al., 2004; Prisacari & Danielson, 2017). Prisacari and Danielson (2017) employed a between-subjects comparison by presenting undergraduate chemistry learners with three chemistry quizzes comprised of differing difficulty (definition, algorithmic, and conceptual). They reported no effect of test mode on either test score or subjective experiences of cognitive load. However, given the variation associated with the experience of intrinsic load at an individual level, the use of within-subjects comparisons is consistent with recommendations made by previous CLT research (Brünken et al., 2002; Debue & Van De Leemput, 2014; Prisacari & Danielson, 2017).

Many studies report that CBT places younger learners at a disadvantage. For example, a study conducted by Bennett et al. (2008) with 1970 eighth grade students reported superior performance on paper over computer-based assessments—a phenomenon also reported in other research focused on primary school-aged children (Chu, 2014; Logan, 2015). According to Endres et al. (2015), one factor that accounts for this difference appears to be WMC. Carpenter and Alloway (2019) conducted working memory assessments in both paper and computer-based modalities with 1339 British children aged 4 to 11 years. They reported poorer scores when

working memory assessments were taken on computers. These findings are consistent with evidence to suggest that CBT environments interfere with students' metacognitive processes (Ackerman & Lauterman, 2012; Noyes & Garland, 2003; Noyes et al., 2004; Sidi et al., 2017). These changes may be reflected in subtle changes in test behaviour, such as the increased use of scratch paper for questions imposing a high IL (Prisacari & Danielson, 2017). Working memory remains a central principle of CLT because the degree to which a learner's WMC is exceeded depends largely on the amount of pre-existing knowledge they hold in LTM.

However, there may also be extraneous factors that consume WM resources independent of prior knowledge. For instance, it has also been observed that the repeated measurement of cognitive load throughout an assessment task may itself impose an extraneous load on learners which, in particular, may disadvantage students with low WMC (Anmarkrud et al., 2019). Therefore, the inclusion of WMC as a covariate in analyses of cognitive load may be a useful way controlling for differences in working memory demands that arise due to the measurement of cognitive load rather than due to addressing requirements of a test question (Anmarkrud et al., 2019; De Jong, 2010). Such measures may provide additional rigour to an analysis because they provide a measure of the "system capacity" (De Jong, 2010, p. 123) that may be influenced by factors other than prior knowledge.

Additionally, if computer-based mediums add a working memory load to learning and assessment tasks otherwise absent in paper-based tasks, it is reasonable to assume that switching one's attention between computer and paper mediums comes at a cost (Collette & Van der Linden, 2002; Ophir et al., 2009; St Clair-Thompson & Gathercole, 2006). For instance, this is likely to occur when completing a test on a computer and relying on scratch paper to support one's calculations. This cost manifests in an increase of the overall cognitive load experienced by a learner and has implications for the way in which students engage with supportive behaviours such as the use of scratch paper during an assessment. For instance, students may rely less on scratch paper when completing a CBT because the computer-based environment itself imposes excessive extraneous load to allow the student to switch attention from screen to paper.

Present Study

The first goal of this study was to advance Prisacari and Danielson's research (2017) of test mode effects on cognitive load and test taking behaviour by focusing on school-aged students. Importantly, the majority of CLT research has focused on optimising learning through instructional sequences. Given the prevalence of relatively high-stakes testing such as PISA and NAPLAN being delivered online, there is a distinct need for consideration of cognitive load experienced under different testing modes. Secondly, this study prioritised comments raised in past research (Ayres, 2015; Martin, 2014; Mayer, 2005; Skulmowski & Xu, 2021) by measuring CL during assessment tasks designed to possess high face validity to students sitting them. Drawing on the work of De Jong (2010) and Skulmowski and Xu (2021), the present study was designed to employ learning and assessment tasks in which students have a specific

interest and motivation to succeed. These authors identified this focus as a key criterion to enhancing the confidence with which classroom practitioners apply CLT research to their own learning and assessment contexts. This was achieved by giving students tests that were meaningful to learners in the context of their in-school assessment programmes. Furthermore, feedback on the quizzes was central to success in an end-of-topic tests students were scheduled to complete in the days following their participation in this study. Moreover, tests were administered with relatively high ecological validity: in students' regular classes by their normal teacher. This included working with classroom teachers and their learners in authentic learning contexts during regular timetabled classes and using assessment material directly linked to their curriculum and assessment programmes.

In referring to tasks that have authenticity, some authors have highlighted the need for greater use of contextualised learning and assessment tasks (Anmarkrud et al., 2019; De Jong, 2010; Martin, 2014). These recommendations are based on observations that learning and assessment tasks that do not appear authentic (i.e., which have low face validity to the students sitting them) are likely to be less reflective of genuine student classroom behaviour (Martin, 2014; Skulmowski & Xu, 2021). Here, De Jong (2010) highlights that a number of cognitive load studies rely on tasks with artificial time constraints or which sample participants who have no specific interest in the domain being assessed (e.g., psychology students learning economics, or biology content—for example, see Park et al. (2015), Schmeck et al. (2015), Van Gog et al. (2012), and Korbach et al. (2018)). The authors note that this may become problematic when drawing generalised conclusions about real-world applications of the underlying theory, particularly from the perspective of classroom practitioners. Therefore, such research ought to prioritise *learning* characteristics by relying on tasks that occur in authentic learning contexts (as defined by De Jong (2010)) with activities that have meaning for students' learning (e.g., tasks that bear direct consequences for students' in-school assessment programmes).

Based on previous findings (Bennett et al., 2008; Carpenter & Alloway, 2019; Chu, 2014; Logan, 2015), it was hypothesised that students would perform better on paper-based tests compared to computer-based tests (H1). Secondly, students would experience increased cognitive load during computer tasks, and the difference between computer and paper-based tests would be greatest for questions of high element interactivity (i.e., difficult questions)—H2. Thirdly, test mode effects would be absent when controlling for individual differences in working memory capacity (H3). Finally, students would demonstrate greater use of scratch paper during paper-based tests compared to computer (H4).

Methods

Participants

The present study was conducted with 263 grade 9 science students at two independent (nongovernment) coeducational high schools in Perth, Western Australia (Index of Cultural and Socio-Economic Advantage (ICSEA)=1158 [97th

percentile] and 1118 [90th percentile] respectively). Data were collected on three occasions during the 2021 and 2022 academic years.

Across both participating schools, there were 465 year 9 students (aged 13–14) enrolled in three cohorts during the data collection. After removing (1) students due to absences, (2) those unable to complete all three required tasks, (3) those who chose not to participate, and (4) those who experienced technical difficulties during data collection, a total of 263 students participated in this study ($M = 140$, $F = 109$, other = 14). Within this sample, there were 14 individual classes taught by 7 different teachers. Given that formative assessment prior to end-of-unit assessment tests is common practice in classrooms, an opt-out consent procedure was implemented. Students were informed that data would be collected in the form of revision quizzes during one regular timetabled science lesson, which served as students' formative assessment, in the week prior to their summative end-of-topic test. This study was approved by the first author's University Human Research Ethics Committee.

Design

A counterbalanced repeated measures design was employed, ensuring students completed all three components (computer and paper quizzes, and working memory capacity test) in a single, regular timetabled science lesson, with all testing procedures administered by students' regular classroom teacher. Approximately half of participants (organised by class groups) completed the paper-based test followed by the computer-based test ($N = 127$) and with remaining participants completing the computer-based test followed by the paper-based test ($N = 136$).

Material

The material used in this study included equivalent versions of a paper-based and computer-based quiz, as well as an online working memory capacity (reverse digit span) test. These materials were written to be commensurate in difficulty and content with the Western Australian Curriculum. The Ohm's Law calculation questions designed for this study were based on common revision questions sourced from widely available student book materials.

Quiz Question Design

Each version of the revision quizzes consisted of eight electrical circuit questions. To ensure equivalency between modalities, eight questions were written and validated for difficulty and element interactivity—in collaboration with four expert teachers (who were colleagues of the lead researcher). The eight questions (including circuit diagrams, question prompts and suggested working) were randomised and presented to the expert teachers who were asked to categorise the questions into two groups of high and low difficulty (based on their perception

of element interactivity). This process resulted in four questions of low difficulty (low element interactivity) intended to place learners under low IL and four questions of high difficulty (high element interactivity) intended to place learners under high IL. The categorisations from the four expert teachers returned 100% agreement on six of the eight questions, with the remaining two questions returning 75% agreement. Further discussion was held with the teachers to clarify discrepancies for these two final questions. This resulted in a set of descriptive criteria that was used for the final categorisation of questions, shown in Table 1.

Following this, each question was duplicated and modified to create an equivalent version of the original test for use in during CBT. The duplicate test contained circuits whose components had a slightly different arrangement and whose values (e.g., assigned voltage, current, or resistance) differed in magnitude. The wording of questions in each mode was identical (see Table 2).

The electrical circuits were drawn and designed using Circuit Diagram online software (Circuit Diagram Editor, 2022), and all images were saved as 370×230 jpeg files for use in both CBT and PBT modalities. Finally, the expert teacher categorisations were validated using the protocol reported by Prisacari and Danielson (2017), by establishing that objective ratings of question difficulty and student performance were negatively correlated.

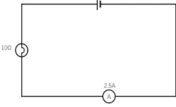
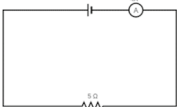
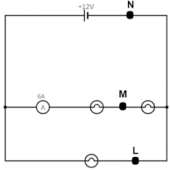
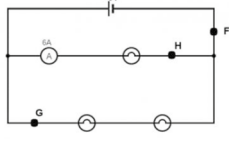
Computer-Based Quiz

The computer-based quiz was designed and distributed using Qualtrics XM software (Qualtrics, 2020). Participants were initially presented with a brief set of written instructions, followed by the eight questions presented in random order. Following each electrical circuit question, students were required to write their answers into a blank text box, before giving two subjective CL ratings using the sliding selector, shown in Fig. 2. Finally, at the bottom of the screen, students were able to navigate between questions freely, regardless of having provided a response or not, by clicking on appropriate “forward” and “back” icons. During CBT, students were instructed to use the back page of their PBT quiz for scratch paper. The Ohm’s Law formula ($V=IR$) was printed at the top of this page because it was assumed that when students needed to refer to this, they would likely also be relying on use of

Table 1 Criteria for categorising easy vs difficult questions

Easy (low EI) question criteria	Difficult (high EI) question criteria
<ul style="list-style-type: none"> • No calculation in some instances (only interpretation of data provided in circuit) • No algebraic transformation of formula • Series circuits • No more than two elements in a circuit • All elements added to a circuit are identical (all bulbs) • At most a simple addition operation is required prior to application of formula 	<ul style="list-style-type: none"> • Algebraic rearrangement of formula • Parallel circuits • Series circuit with 2 or more elements whose values (e.g., resistance) must be derived from the total value (e.g., resistance) of the circuit through division after the formula for Ohm’s law has been applied

Table 2 Examples of equivalent easy vs difficult questions presented in computer and paper modes

Paper-based question	Equivalent computer-based question	Suggested working
		<p>Calculate the voltage across the resistor in this circuit.</p> <p>Voltage = Current x Resistance</p> <p>Voltage = 4 x 5</p> <p>Voltage = 20 V</p>
		<p>Predict the current flowing through point F in the circuit above, assuming the light bulbs are of equal resistance.</p> <p>Resistance in bottom circuit</p> <p>2 is 2 x resistance in top circuit = 6 A</p> <p>So, current in bottom circuit = 1/2 current in bottom circuit = 3A</p> <p>Current at point F = 3 + 9 = 12 amps</p>

This example includes two circuit-based electricity problems of low element interactivity (low intrinsic load) in the top row and high element interactivity (high intrinsic load) in the bottom row. In the first example (top row), students were provided with the formula relating voltage, current, and resistance. Students needed to identify the correct variables and calculate the final voltage. The second question (bottom row) is high in element interactivity for two reasons. Firstly, the circuit has many more components to consider simultaneously. Secondly, students were required to devise the algebraic relationship between the resistance in each half of the circuit themselves in order to answer the question correctly, placing them under relatively high cognitive (intrinsic load)

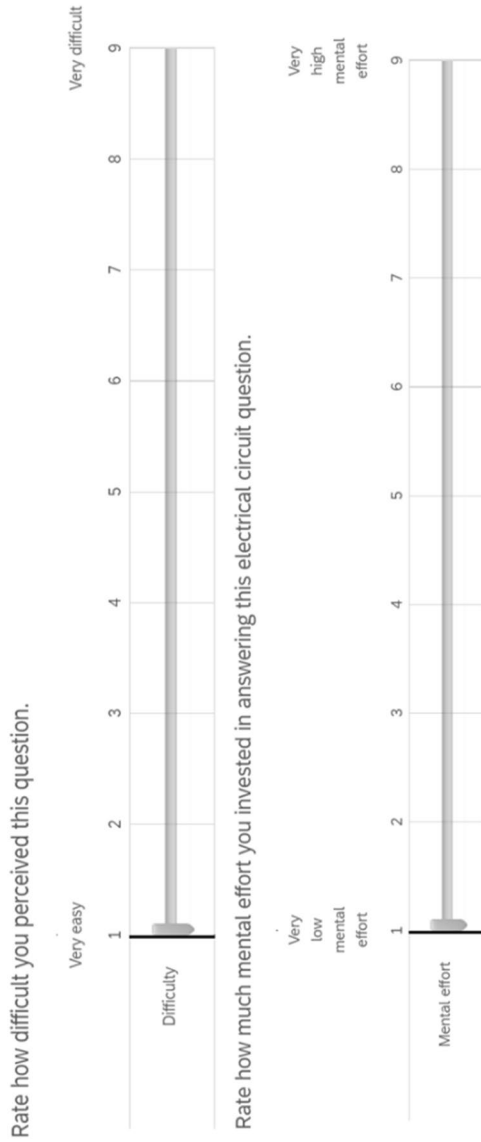


Fig. 2 The subjective cognitive load scales that followed each CBT question

scratch paper to support their working. Providing the formula here therefore limited the number of times students would be required to switch between paper and computer media.

Paper-Based Quiz

The paper-based quiz was created using MS Word (Arial, 22-point font) and contained a cover page featuring instructions and the relevant University logo in the page header, followed by the eight questions (one on each page). To limit order effects, four versions of the PBT were created, each with questions in a randomised order. Each question (including the back page which was provided as scratch paper for use during CBT) was presented with the Ohm's Law formula ($V=IR$) at the top of the page in bold font and electrical circuit diagram aligned to the top left of the page (to allow scratch paper working space in the adjacent area). Under each circuit diagram, students were presented with an imperative prompt (e.g., "Calculate the resistance of the light bulb in the circuit above") and then prompted to provide their answer on a blank line. Under this, students were provided a replica image of the 9-point difficulty and mental effort scales shown in Fig. 2, which they marked in pen.

Working Memory Capacity Assessment

This study used an adaptive reverse digit span protocol for measuring WMC. The digit span WMC test is a commonly used procedure for measuring working memory in CLT research (Carpenter & Alloway, 2019; Chen et al., 2018; Mayes et al., 2001) and has been an integral component of Weschler Intelligence Scales. The reverse digit span protocol presents participants with single digits one at a time, with participants required to recall the numbers in reverse order of presentation, thus placing them under relatively higher working memory load than the forward digit span protocol, and serving as a more time-efficient measure of WMC (Maerlender et al., 2004; Wilde et al., 2004).

The WMC was constructed using Qualtrics XM software (Qualtrics, 2020) using a procedure consistent with descriptions provided by Kessels et al. (2008) and Wilde et al. (2004). Noteworthy, an adaptive procedure was incorporated for time efficiency, allowing the completing of CBT, PBT, and WMC tasks to be completed within a single lesson (roughly 55 min). Participants were initially presented with a brief set of written instructions on the screen. Each trial was initiated with a "+" priming cue for one second. For each trial, single digits (0–9) were presented at random for one second, each followed by a blank (white) spacer for 0.5 s. The test began with the presentation of a number string of two digits. The number string increased by one digit each time participants correctly recalled the sequence in reverse order. Following a false response, the subsequent trial presented students with a number string of identical length. The test continued until participants returned two consecutive incorrect trials. WMC was deemed to be the last digit string length that was correctly recalled. Immediately following the termination of the digit span test, students were prompted to enter information about their gender and preference for testing mode.

Measuring Cognitive Load

Cognitive load has historically been measured using three broad techniques: indirect measures (such as subjective rating scales), secondary task measures, and physiological measures (Prisacari & Danielson, 2017; Sweller et al., 2011). Whilst a detailed consideration of each category is beyond the scope of this article, the focus of this study is subjective participant ratings of CL based on a 9-point rating scale (Bratfisch et al., 1972; Paas et al., 1994). The use of subjective scales in CLT research suggests they are sensitive to task complexity and have relatively high levels of reliability (Ayres, 2006; Klepsch et al., 2017; Paas et al., 1994). Ayres (2006) employed a 7-point CL scale following a series of mathematical bracket expansion tasks. Whilst it was concluded that subjective CL measures were sensitive to subtle changes in the experience of CL that were not reflected in learners' rates of error, it has been suggested that subjective scales may be less valid when used with low-level learners who have poorer schema formation and thus poorer grasp of task complexities (Naismith et al., 2015).

The present study employed two 9-point scales of perceived difficulty and mental effort (Fig. 2), which is more commonly used than the equivalent 7-point Paas scale (Sweller & Paas, 2017). Following each question, provided ratings of difficulty and effort and these scores were averaged for easy (low element interactivity) and difficult (high element interactivity) questions.

These 7-point and 9-point scales have reported Cronbach's alpha scores of between 0.6 and 0.9 (Ayres, 2006; Klepsch et al., 2017; Paas et al., 1994). Internal consistency measures for the present study ranged between 0.91 and 0.93. Taking measurements following every question provides a sensitive, valid, and reliable measure of item-level variations in CL (Klepsch et al., 2017; Korbach et al., 2018; Prisacari & Danielson, 2017; Schmeck et al., 2015; Sweller & Paas, 2017). Additionally, compared to physiological measures and secondary task measures, the subjective measures of CL provide the least intrusive, time efficient, and most ecologically valid way to detect item-level changes in cognitive load (Korbach et al., 2018; Prisacari & Danielson, 2017; Raaijmakers et al., 2017).

Test-Taking Behaviour (Scratch Paper Use)

Prisacari and Danielson (2017) operationalised test-taking behaviour with a simple count of whether students used scratch paper or not. This was achieved by scoring each question as a 0 (student did not use scratch paper) or a 1 (student did use scratch paper). However, this provided a rather crude measure of scratch paper use. The present study addressed recommendations put forward by Prisacari and Danielson (2017) to consider *how much* students write on scratch paper (e.g., the number of "moves" used on scratch paper). For both versions (CBT and PBT), students were provided scratch paper space. In the PBT, students were provided with working space to the right-hand side of each circuit diagram. For the CBT, students were instructed to use the back page of their PBT, which was provided as a blank page and titled "Please use this page for working out during your computer-based revision quiz." To account for testing

mode differences in scratch paper use during calculations in each quiz, the numbers of working out steps taken by students were counted. Writing or rearranging a formula, identifying variables in a question, or adding variables together were each counted as one step. Any responses that could not be directly linked to a question or could not be interpreted were ignored. To ensure reliability of this procedure, the lead researcher and an independent reviewer initially moderated 10 papers, reaching a consensus on all discrepancies following a brief discussion. Following this, an additional 10 papers were co-moderated, with both reviewers achieving agreement on 95% of all questions. Finally, the remaining papers were assessed for scratch paper use in both modalities by the independent reviewer.

Procedure

Prior to data collection, participating classroom teachers were briefed by the lead researcher, receiving a complete set of PBT materials, PBT answer keys, and scripted instructions and URLs for the CBT and WMC tests. Scripted instructions read by teachers were designed such that participants were presented with a brief learner training explanation of cognitive load and element interactivity as this has been shown to increase learners' ability to discriminate between questions of high and low cognitive load (Klepsch et al., 2017). Students were first instructed to generate their own unique de-identified code. This was recorded at the start of both test modalities and the WMC test for data matching purposes. Students were given 15 min for each version of the test. Following this, students were given 10 min to complete the WMC test. Finally, students were provided with a marking key corresponding to their paper-based quiz version (feedback was not provided on answers given to CBT questions due to time constraints and the fact that questions in each testing mode relied on identical processes) and instructed to self-assess their own paper quiz for formative feedback using a different coloured pen, under teacher supervision, and provided with assistance for questions of concern. The total time for data collection was approximately 50 min.

Data Analysis and Modelling Approach

Hierarchical linear models were used for the main analysis using Jamovi (The jamovi project, 2022) with random effects included in each model to capture the study design (i.e., the repeated measures of students, nested within classrooms, nested within teachers, and nested within cohort). Rather than using HLM to engage in a model selection process, the technique was used to test for main effects and interactions pertaining to hypotheses 1–4.

Testing Mode Effects on Test Score

To address hypothesis 1, a hierarchical linear model was fitted to the response variable test scores (model 1). The fixed effects captured the research objectives (i.e., to assess the effect of assessment mode (CBT vs PBT) and whether that difference

depended on question difficulty). To control for individual differences in working memory, WMC was added as a covariate to model 1a.

Testing Mode Effects on Cognitive Load

Because cognitive load was operationalised by both students' ratings of perceived difficulty of questions and mental effort invested, two hierarchical linear models were fitted to address hypothesis 2. The first was fitted with perceived difficulty as the response variable (model 2). Like model 1, the fixed effects captured the research objectives (i.e., to assess the effect of assessment mode and whether that difference depended on question difficulty).

For the second model assessment mode effects on cognitive load, a hierarchical linear model was fitted, with mental effort as the response variable (model 3), and fixed effects capturing the effect of testing mode on cognitive load, and whether the difference depended on question difficulty. To control for individual differences in working memory, WMC was added as a covariate to models 2a and 2b.

Testing Mode Differences in Test-Taking Behaviour

To address hypothesis 4, a hierarchical linear model was fitted to the response variable scratch paper use, with the same fixed and random effect structure as the previous models (model 4).

Results

For all statistical tests employed in the following analysis, an alpha level of 0.05 was used (in some instances, a Bonferroni correction was applied). All significance values for mean differences are two tailed. Exact p values have been reported except for very small values, in which was <0.001 is used.

Validation of Test Materials

To verify the subjective categorisation of quiz questions according to difficulty (i.e., easy vs difficult), the protocol reported by Prisacari and Danielson (2017) was used. This resulted in a significant, strong negative correlation between subjective difficulty ratings provided by students and the objective difficulty determined by the proportion of students correctly answering each question on paper ($r(6) = -0.966$, $p < 0.001$) and computer ($r(6) = -0.875$, $p = 0.004$). Furthermore, to ensure that equivalency of all four paper versions of the quiz and to mitigate for potential order effects due to the limited number of possible permutations of question order, a one-way ANOVA was conducted. There were no significant differences in test scores for the four versions of the PBT ($F(3, 259) = 1.31$, $p = 0.273$).

Nested Data Structure

To validate the nested structure of the data used in this analysis, intraclass correlations were calculated for the four main models used to test hypotheses H1 to H4, shown in Table 3. From this analysis, between 51 and 73% of variance in the data was attributed to variability between students. Between 7 and 19% of the variance was attributable to differences between class groups, and between 2 and 9% of the variance was attributable to differences between cohorts. Interestingly, negligible variance in the data was attributable to differences between teachers.

Testing Mode Effects on Test Score

This model indicated that students performed better on easy CBT questions ($M=2.92$, 95% CI=2.24, 3.61) compared to easy PBT questions ($M=2.65$, 95% CI=1.96, 3.34), whilst students performed better on difficult PBT questions ($M=1.50$, 95% CI=0.73, 2.26) compared to difficult CBT questions ($M=1.03$, 95% CI=0.28, 1.79). Out of a total possible score of 4, the difference between modalities for easy questions (0.27) represents approximately 7%, whilst the difference between modalities for difficult questions (0.46) represents 11.5%. As shown in Fig. 3, there was a significant main effect for mode ($F(1,767.18)=4.07$, $p=0.044$); difficulty ($F(1,7.82)=570.92$, $p<0.001$); and also for the mode*difficulty interaction ($F(1,759.66)=63.14$, $p<0.001$). Bonferroni adjusted post-hoc comparisons indicated that the differences between mode for easy questions were significant ($t(763.1)=4.17$, $p<0.001$). Testing mode differences for difficult questions were also significant ($t(763.2)=-7.03$, $p<0.001$). See Table 4 for full model results.

Testing Mode Effects on Cognitive Load

Students rated paper questions as more difficult ($M=4.17$, 95% CI=3.46, 4.88) compared to computer questions ($M=3.97$, 95% CI=3.26, 4.68). As shown in Fig. 4, there was a significant main effect for mode ($F(1,760)=11.66$, $p<0.001$); for difficulty ($F(1, 756)=162.63$, $p<0.001$); but not for the Testing mode*difficulty interaction ($F(1,756)=3.80$, $p=0.052$). Bonferroni-adjusted post-hoc comparisons

Table 3 Intraclass correlations for the four main models analysed

Model	Outcome variable	Students	Class groups	Teacher	Cohort
1	Test scores	0.51	0.07	<0.001	0.09
1a (with WMC)	Test scores	0.49	0.05	0.00	0.06
2	Perceived difficulty	0.73	0.16	0.00	0.02
2a (with WMC)	Perceived difficulty	0.73	0.16	0.00	0.01
3	Mental effort	0.67	0.19	0.00	0.02
3a (with WMC)	Mental effort	0.67	0.17	0.00	0.01
4	Scratch paper	0.64	0.08	0.00	0.07

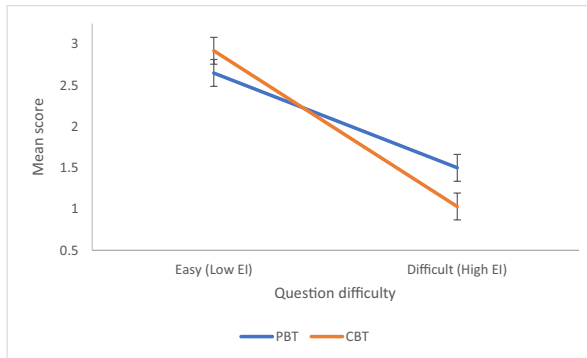


Fig. 3 Interaction between testing mode and question difficulty (element interactivity). Note. Error bars show standard error of estimated marginal means

Table 4 Effect of testing mode on test score (models 1 and 1a)

	Model 1			Model 1a (with WMC)		
	<i>B</i>	95% CI	<i>p</i>	<i>B</i>	95% CI	<i>p</i>
Fixed effects						
(Intercept)	2.04	[1.72–2.35]	0.006	2.04	[1.80–2.23]	0.004
Test mode	0.09	[0.002–0.19]	0.045	0.09	[–0.002–0.18]	0.06
Question difficulty	–1.53	[–1.62 to –1.44]	<0.001	–1.52	[–1.61–1.43]	<0.001
Mode × difficulty	0.74	[0.55–0.92]	<0.001	0.74	[0.56–0.92]	<0.001
WMC				0.12	[0.07–0.17]	<0.001
Random effects	Variance			Variance		
σ^2	0.56			0.75		
Participant ID	0.57			0.73		
Class group	0.04			0.17		
Teacher	0.000			0.000		
Cohort	0.05			0.18		
Marginal R^2	0.33			0.75		

indicated that the testing mode differences for easy questions were not significant ($t(752) = -1.05, p = 1.00$). Testing mode differences for difficulty questions were significant ($t(752.1) = -3.79, p < 0.001$). See Table 5 for full model results.

Students invested more effort in paper questions ($M = 4.08, 95\% \text{ CI} = 3.19, 4.96$) compared to computer questions ($M = 3.71, 95\% = 2.83, 4.59$). Bonferroni-adjusted post-hoc comparisons indicated a significant difference between modalities for easy questions ($t(752.7) = -5.19, p < 0.001$) and also difficult questions ($t(752.8) = -2.73, p = 0.039$). As shown in Fig. 5, this model resulted in a significant main effect for mode ($F(1,760.7) = 31.14, p < 0.001$) and difficulty ($F(1,37.6) = 204.85, p < 0.001$), but not for the testing mode*difficulty interaction ($F(1,755.6) = 3.03, p = 0.08$). See Table 6 for full model results.

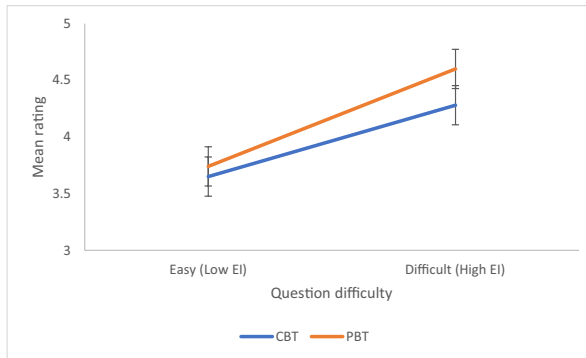


Fig. 4 Perceived difficulty of questions by testing mode and question difficulty (element interactivity). Note. Error bars show standard error of estimated marginal means

Table 5 Effect of testing mode on perceived difficulty (models 2 and 2a)

	Model 2			Model 2a (with WMC)		
Fixed effects	<i>B</i>	95% CI	<i>p</i>	<i>B</i>	95% CI	<i>p</i>
(Intercept)	4.07	[3.75–4.93]	0.002	4.06	[3.76–4.37]	0.002
Test mode	0.20	[0.09–0.31]	<0.001	0.20	[0.09–0.32]	<0.001
Question difficulty	0.75	[0.63–0.86]	<0.001	0.75	[0.63–0.86]	<0.001
Mode × difficulty	0.23	[–0.001–0.46]	0.052	0.22	[–0.01–0.45]	0.06
WMC				–0.04	[–0.14–0.07]	0.51
Random effects	Variance			Variance		
σ^2	0.89			0.88		
Participant ID	2.37			2.38		
Class group	0.17			0.16		
Teacher	0.00			0.00		
Cohort	0.01			0.007		
Marginal R^2	0.04			0.04		

Effect of Working Memory on Test Score and Cognitive Load

This model (model 2a, Table 5) resulted in a nonsignificant effect of WMC on perceived difficulty ($F(1,250)=0.45$, $p=0.51$); significant effects for mode ($F(1,757)=11.92$, $p<0.001$); difficulty ($F(1,753)=161.0$, $p<0.001$); and a nonsignificant testing mode*difficulty interaction ($F(1,753)=4.49$, $p=0.06$). Secondly, when WMC was added to the model fitted with mental effort as the outcome variable (model 3a, Table 6), this resulted in a nonsignificant effect of WMC on mental effort ($F(1,255)=2.03$, $p=0.16$); significant effects for testing mode ($F(1,758)=31.66$,

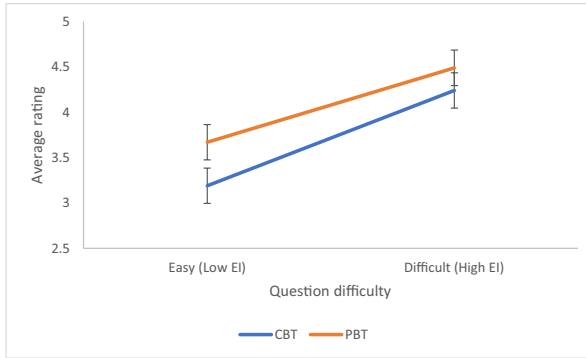


Fig. 5 Mental effort invested in questions according to testing mode and question difficulty. Note. Error bars show standard error of estimated marginal means

Table 6 Effect of testing mode on mental effort (models 3 and 3a)

	Model 3			Model 3a (with WMC)		
Fixed effects	<i>B</i>	95% CI	<i>p</i>	<i>B</i>	95% CI	<i>p</i>
(Intercept)	3.90	[3.53–4.27]	0.003	3.89	[3.56–4.21]	0.003
Test mode	0.36	[0.24–0.49]	<0.001	0.37	[0.24–0.49]	<0.001
Question difficulty	0.93	[0.80–1.06]	<0.001	0.93	[0.80–1.06]	<0.001
Mode × difficulty	−0.22	[−0.48–0.03]	0.08	−0.24	[−0.49–0.01]	0.07
WMC				−0.07	[−0.18–0.03]	0.16
Random effects	Variance			Variance		
σ^2	1.06			1.05		
Participant ID	2.13			2.14		
Class group	0.25			0.22		
Teacher	0.00			0.00		
Cohort	0.02			0.007		
Marginal R^2	0.07			0.07		

$p < 0.001$); difficulty ($F(1,752) = 207.97, p < 0.001$); and a nonsignificant testing mode*difficulty interaction ($F(1,752) = 3.40, p = 0.07$). Finally, controlling for individual differences in WMC on test scores (model 1a, Table 4) resulted in a nonsignificant effect of mode ($F(1, 768) = 3.64, p = 0.057$); significant effects for difficulty ($F(1,760) = 1073.95, p < 0.001$); WMC ($F(1,252) = 19.45, p < 0.001$); and also the testing mode*difficulty interaction ($F(1,760) = 63.64, p < 0.001$).

Testing Mode Effects on Test Behaviour (Scratch Paper Use)

Students used scratch paper more during PBT ($M = 1.91, 95\% \text{ CI} = 0.30, 3.52$) compared to CBT ($M = 1.26, 95\% \text{ CI} = -0.35, 2.87$). Bonferroni-adjusted post-hoc

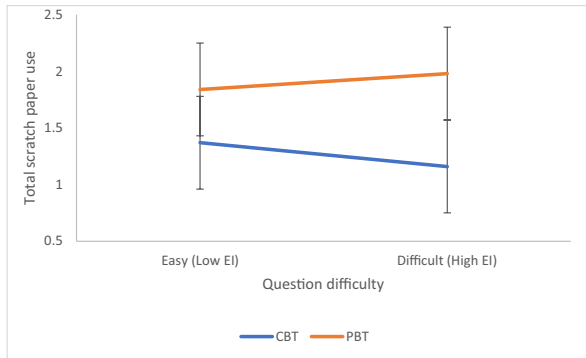


Fig. 6 Scratch paper use by testing mode and question difficulty. Note. Error bars show standard error of estimated marginal means

Table 7 Effect of testing mode on scratch paper use

		Model 4		
		<i>B</i>	95% CI	<i>p</i>
Fixed effects	(Intercept)	1.59	[0.81–2.36]	0.06
	Test mode	0.65	[0.40–0.90]	<0.001
	Question difficulty	−0.04	[−0.28–0.21]	0.78
	Mode × difficulty	0.34	[−0.15–0.84]	0.17
Random effects	Variance			
	σ^2	4.07		
	Participant ID	7.10		
	Class group	0.35		
	Teacher	0.00		
	Cohort	0.30		
	Marginal R^2	0.009		

comparisons revealed a significant difference in scratch paper use for both easy questions ($t(766) = -2.69$, $p = 0.044$) and difficult questions ($t(766) = -4.61$, $p < 0.001$). Of key interest is that not only did students use scratch paper less during CBT, but their use of scratch paper decreased as question difficulty increased. As shown in Fig. 6, this model resulted in a significant main effect for mode ($F(1,767) = 26.61$, $p < 0.001$), a non-significant effect for difficulty ($F(1,766) = 0.08$, $p = 0.78$), and nonsignificant testing mode*difficulty interaction ($F(1,766) = 1.86$, $p = 0.17$). See Table 7 for full model results.

Discussion

The present study addressed whether the effects of manipulating extraneous cognitive load through test mode year 9 students' experiences of cognitive load and test behaviour and whether these effects were dependent on increases in intrinsic

cognitive load (determined by question difficulty) and individual differences in working memory capacity. The results from this study provide support for hypothesis 1 (that students would perform better on PBT questions). The results revealed an inverse relationship of that predicted by hypothesis 2 (that students would report greater cognitive load for CBT questions). The results partially supported hypothesis 3 (that test mode effects would be absent when controlling for differences in WMC), because this was only true for models with test score as the outcome variable. It was not the case for models with measures of cognitive load as the outcome variable. Finally, the present results support hypothesis 4 (that scratch paper use would be greatest for PBT). Surprisingly, scratch paper use decreased during CBT as question difficulty increased.

Methodologically, the present study sought to address recommendations in previous research highlighting the need for greater use of repeated measures designs, in authentic learning contexts with high face and ecological validity (Anmarkrud et al., 2019; De Jong, 2010; Skulmowski & Xu, 2021). Of the studies that have employed a within-subjects comparison of performance in paper and computer modality, most indicate that students perform better on paper than computer in both test scores (Carpenter & Alloway, 2019; Hardre et al., 2007) and memory formation processes (Noyes & Garland, 2003). The findings based on present methodology support these claims and therefore make a substantial contribution to CL and testing mode research.

The present findings are consistent with previous studies which have shown that primary and secondary school-aged students achieve better scores when taking their assessments on paper, and that testing mode effects depend on both intrinsic and extraneous loads imposed learning and assessment tasks (Bennett et al., 2008; Carpenter & Alloway, 2019; Chu, 2014; Logan, 2015). This suggests that school-aged students are increasingly disadvantaged when answering questions on a computer as the task difficulty increases. Out of a total possible score of 4, the difference between modalities for easy questions (0.27) represents approximately a difference of 7%, whilst the difference between modalities for difficult questions (0.46) represents a difference of 11.5%. This contrasts with previous research suggesting that testing mode has no effect on test outcomes or the experience of cognitive load (Prisacari & Danielson, 2017).

Contrary to expectations however, students performed better on easy CBT questions compared to PBT. One plausible explanation for this lies in the additive nature of intrinsic and extraneous sources of load. In comparing differences between testing modalities, the present study controlled for variations in intrinsic load by using equivalent forms of test questions. CLT posits that learning environments are most effective with EL is minimised and IL is optimised within the limits of students' WMC (Van Merriënboer et al., 2006). So, where a student is under relatively low IL, factors such as motivation to engage or differences in attention control are likely to influence the extent to which a student answers question correctly (Schnotz, 2010; Schnotz & Kürschner, 2007). For instance, school-aged students have been shown to read on-screen information more quickly, yet they also overestimate their comprehension of these on-screen texts (Ronconi et al., 2022). Therefore, it is plausible that for easy questions, the higher CBT scores on easy questions can be accounted for because on-screen tasks are more engaging, or demand greater attentional resources without placing students under excessive cognitive load. In other words, the CBT

mode may be sufficiently increasing germane processing under low intrinsic load (De Jong, 2010; Skulmowski & Xu, 2021), thus accounting for the higher test scores. This is consistent with research by Likourezos et al. (2019) who presented expert and novice learners with materials of increasing variability and found that knowledgeable learners perform more poorly on tasks that induce low levels of load. This supports the assertion that task outcomes are enhanced when intrinsic load is optimised within the limits of total cognitive load, especially as learner expertise increases. Conversely, as question difficulty (i.e., intrinsic load) increases, students are placed under greater total load. If one accepts the assumption that computer-based testing environments add extraneous load, the present findings are consistent with the CLT. This is because once IL (i.e., question difficulty) exceeds a student's working memory capacity, the added EL associated with the computer-based testing environment becomes deleterious to performance, and thus test scores decrease. Together, this finding highlights not only the fact the testing mode affects test performance in school-aged students, but that understanding this effect requires consideration of other mediating factors such as cognitive load.

Test Mode Effects on Cognitive Load

One key aspect of this study was the effect of test mode on the experience of cognitive load, measured by mental effort and perceived difficulty of questions. The results indicated that students perceived paper-based tests as more difficult and invested more effort in these questions. There was a marked difference between the two modalities in the mental effort reported for easy questions, with students investing much greater effort in easy PBT questions, despite performing better on computer for these questions. Given the previous evidence to support the increased demand placed on working memory capacity associated with the extraneous load imposed by computer-based tests (Carpenter & Alloway, 2019), one might expect that computer-based modalities would lead to an increase in students' reports of cognitive load. Past research has shown that when students are under low levels of intrinsic load, the deleterious effect of extraneous load on performance decreases and, in some instances, reverses (Sweller & Chandler, 1994). For instance, Park et al. (2011) measured the effect of extraneous distracting information on learning high school science content using a multimedia environment. Their results indicated that extraneous load had beneficial effects on student learning under low levels of additional cognitive load. However, given that EL and IL have been shown to operate in an additive manner (Jiang & Kalyuga, 2020; Kalyuga, 2011; Skulmowski & Xu, 2021; Sweller et al., 2011), we offer three interpretations to account for the present results.

Firstly, whilst both participating schools employ a 1:1 device policy (whereby all students are equipped with their own laptop or iPad), both schools provide summative assessment to students on paper. Therefore, a logical conclusion must address the fact that students in both schools were very familiar with computer-based environments and more likely to report relatively low levels of load associated with their use. However, the participating students are also more familiar (and therefore more

confident) in answering assessment questions on paper. This increased familiarity with paper-based modality for assessment tasks may explain the increase in effort invested in answering paper-based questions. This view of students' experience of CL is consistent with previous research that underscores the relationship between the experience of cognitive load and one's motivation to engage with it (Jiang & Kalyuga, 2020; Kalyuga & Singh, 2016; Schnotz & Kürschner, 2007).

A second explanation to account for the increase in CL during PBT may lie in the measurement of CL itself. The scales for difficulty and effort used in the present study refer specifically to the nature of the question being answered (e.g., "rate how difficult you perceived this question") with no explicit reference to the extraneous difficulty or effort associated with answering the question in a specific modality (e.g., "rate how difficult it was to answer this question on paper"). This may indicate that the subjective scales used were sensitive to perceptions of differences in *intrinsic* load, but less sensitive to differences in *extraneous* load imposed by testing mode, or that students were placed under insufficient load during easy questions to provide accurate estimates using the Paas scale (Sweller, 2018).

A third explanation of the present results suggests that paper-based tests may have greater face validity for students, and therefore induce more performance-conducive behaviours in learners. This is consistent with findings that when presented with information on paper, students are better able to reconcile pieces of separate information compared to information presented on computer, as measured objectively by eye tracking techniques (Latini et al., 2020). Because improved test scores were concomitant with an increase in reported extraneous load, an important question that research must now consider is what the optimal theoretical and empirical threshold of CL might be in a given context. Additionally, testing formats that induce greater effort (rather than those that are the easiest to complete, administer or mark) will align with those that have the greatest face validity for learners and are likely to produce better academic outcomes.

Importantly, although the design of the present study ensured that intrinsic load was kept constant between modalities, students reported higher levels of load under PBT. This may reflect increases in germane processing—or the processes involved in information that is intrinsic to the assessment questions (Sweller et al., 2019). Although no longer conceptualised as contributing to the overall experience of cognitive load, germane processes have been associated with motivational factors that increase engagement with a task (Skulmowski & Xu, 2021). For instance, it has been suggested that the costs associated with greater extraneous load induced by CBT modes mediated by individual student characteristics such as motivation, affect, or physiological responses (Choi et al., 2014). In turn, this may contribute to increased germane processing (De Jong, 2010; Skulmowski & Xu, 2021). If students are more motivated to engage with one assessment environment over another, this may lead to them experiencing greater germane processing as they answer questions. Whilst the intrinsic load induced by questions across both CBT and PBT should have remained the same, and whilst the extraneous effect of CBT did not lead to the predicted increase in load, it is possible that students reported greater investment of effort in PBT because they appeared more valid as a form of assessment.

Skulmowski and Xu (2021) highlight that forms of assessment associated with a given extraneous load must be chosen based on the nature of germane processing they stimulate. Therefore, assessment environments must seek to minimise extraneous loading that is not directly linked to germane processing. In other words, CBT environments may have an unavoidable extraneous load cost. However, they may also invoke a beneficial increase in germane processing, leading to better test outcomes. Applying this line of thinking to the present results, as question difficulty increased, the cost of increased in extraneous load derived from CBT may ultimately have exceeded the concomitant benefit associated with greater germane processing invoked by the on-screen environment.

The Role of Working Memory in Cognitive Load and Test Behaviour

Previous research has shown that computer-based tasks interfere with student metacognition, leading to subtle changes in test behaviour (Ackerman & Lauterman, 2012; Mayes et al., 2001; Noyes & Garland, 2003; Noyes et al., 2004; Sidi et al., 2017). Scratch paper use during assessment tasks has previously been associated with more favourable outcomes and is thought to be reflective of working memory processes required to answer questions correctly (Bennett et al., 2008; Prisacari & Danielson, 2017). The present study compared scratch paper use in both modalities and observed significantly more scratch paper use in the PBT, with the biggest difference between the modalities occurring for difficult (high IL) questions. These findings are similar to those of Prisacari and Danielson (2017) who found an increase in scratch paper use on difficult PBT questions. However, the present data reflects a negative trend in scratch paper use as IL increases during CBT. Consistent with the research by Sidi et al. (2017), one explanation is that working memory is sensitive to extraneous load of computer-based tasks in ways that result in changes to underlying metacognitive processes. The present data supports this by showing that these changes in metacognitive processes are observable in differences in test behaviour, such as scratch paper use. Furthermore, the extraneous load imposed by computer-based tasks may not be immediately obvious until learners are placed under increasing IL, which accounts for the decrease in performance on difficult CBT questions.

An alternative explanation for the greater use of scratch paper for paper compared to computer tests is that there may be a cognitive load cost to switching from between computer and paper media (De Jong, 2010). Whilst completing PBT, the scratch paper available to students was adjacent to the question and required minimal shift of attention to be accessed. On the other hand, during CBT, students may be required to shift their gaze from the computer screen to their piece of scratch paper, locate and pick up their pencil and transfer details from the question from their screen onto the page. Indeed, the decrease of scratch paper use as IL increased for question difficulty may reflect the fact that as IL increases, students have fewer working memory resource available and cannot afford the cognitive load cost associated with shifting the details of the question from the screen to paper. Furthermore, it may also be possible that CBT scores were lower for difficult questions because

accessing the scratch paper required to support their working was associated with too great a cognitive load cost.

However, given that scratch paper use is associated with more positive test outcomes, increasing scratch paper use alongside more difficult questions should be encouraged, regardless of the mode in which a test is taken. These results do indicate that school-aged students may benefit from greater learner training and meta-cognitive strategy support when using computer-based assessments because the use of scratch paper in these contexts is associated with better outcomes. However, an important question that remains unaddressed is the effect of offering digital scratch paper space that can be accessed without forcing students to switch between on-screen and paper environments during an assessment.

Outside CL research, there is also evidence that suggests the way learners interact with computers fundamentally changes their cognitive approaches to organising information during learning. For instance, research by Mueller and Oppenheimer (2014) showed that students who take notes on a computer (compared to those who write paper-based notes by hand) tend to write information verbatim, rather than process content more deeply, synthesising it as they work. The outcome of this subtle shift in behaviour is a decrease in performance on conceptual questions that, compared to factual recall questions, impose a greater intrinsic load on students. These findings are consistent with the decrease in scratch paper use and concomitant decrease in performance on difficult CBT questions observed in the present study.

In relation to the role of working memory, this study offered a significant contribution by including working memory capacity measures in its analysis. It was hypothesised that testing mode effects would depend on individual differences in WMC; however, this was only true when considering the mode effects on test scores, but not for the modality*difficulty interaction or either measure of CL. Based on these results, it appears that individual differences in WMC may moderate the effect that testing mode has on test scores, although the effects of adding WMC to the models in this study were small. Of interest is the fact that there was no change in the modality*difficulty effect on test scores when controlling for WMC. One possible reason for this may relate to the sensitivity of the WMC measure used in the present design. If the effect of testing mode is inherently small and the effect of question difficulty is much larger, then using a low-sensitivity measure of WMC is unlikely to produce drastic shifts in modelling.

These results support previous findings indicating that testing mode may inequitably affect school-aged students with lower WMC (Batka & Peterson, 2005; Logan, 2015). Carpenter and Alloway (2019) concluded that test mode effects may result from cognitive load imposed by CBT environments; however, the results from the present study do not support this because the addition of WMC measures rendered no statistical change in test mode effects for models predicting either measure of CL. This finding presents an interesting question for future CL research because despite being founded theoretically on the architecture of human cognition; there was limited support for the causal link between WMC and the experience of cognitive load. However, it is also worth acknowledging that the extent to which a student experiences cognitive overload operates as a function of background knowledge (Sweller, 2020), for which there are no practical precise measures, and so the influence of

WMC as a covariate in CLT analyses may be inseparable from this. Future research should also consider using alternative measures of WMC to assess the validity of this finding.

Limitations and Future Research

The intention of the current study was to explore the differences between paper and computer-based assessments in science classrooms. Data was only collected from students attending schools with relatively high socio-economic advantage. It should be acknowledged that the results from the hierarchical linear models used in the present study suggest that the majority of variance (in excess of 50%) was attributable to differences between individual students, rather than class (7%), teacher (0%), or cohort (9%) factors. However, a wider range of data from students of more diverse backgrounds would be beneficial. Additionally, the present methodology meant it was impossible to control how students were interacting with their devices during CBT. Whilst there is evidence supporting claims of an increase in digital distraction when learning on a computer (Ehrlick, 2014; Flanigan & Titsworth, 2020) and whilst data points that indicated students had spent excessive amounts of time on any CBT task were removed, it is possible that learners were using notes or online resources during the CBT tasks, depending on how the task was proctored by classroom teachers. Future research would benefit from considering different task types that reflect computer literacy, such as qualitative response tasks (e.g., text construction and writing tasks that are dependent on skills such as typing speed). Because computer-based tasks have been shown to increase task completion time (Ronconi et al., 2022), tasks that take longer to complete may exacerbate working memory depletion effects (Chen et al., 2018). This line of inquiry is especially important, given the assumptions behind the rationale for the online delivery of national level testing, such as NAPLAN, for which it is claimed:

While access to computers at home or at school varies, students' performance during the test is likely to depend on how familiar they are with the device they are using for the test, rather than how often they use a computer... ACARA research shows that online writing is similar to handwriting in terms of the quality of writing produced by students at each year level (National Assessment Programme, FAQs, NAPLAN – General, 2016)

The findings from the present study do not support this assumption because familiarity with computers was largely controlled for, yet there were clear differences in both test score and cognitive load experienced due to differences in testing mode. However, an important factor to establish in future research is whether these effects are also observed in extended writing tasks that are commonly found in literacy assessments such as NAPLAN which require students to construct qualitative texts.

This study highlights the need for ongoing educational research in authentic learning contexts. Future research in this area may consider testing mode effects on patterns of learner interaction with assessment material (such as changes in the

length of time to complete assessments, or student sensitivity to distracting information associated with navigational features of computer-based assessment tasks). There has been some exciting work using pupillometry in the measurement of CL which may offer the technology to explore test mode effects on the division of attentional resources. These resources may rely on temporal dimensions of visual searches to deduce how students devote attention during digital learning (Debut & Van De Leemput, 2014; Szulewski et al., 2017).

Secondly, measurement of cognitive load remains a contentious topic in CL research. Subjective scales are not without their criticisms, and their validity may be increased by triangulating data with the use of objective or physiological measures, such as ECG, eye-tracking tools (Ayres et al., 2021; Galy et al., 2012; Solhjoo et al., 2019), and fMRI (Whelan, 2007). A recent review of studies employing physiological and subjective measures by Ayres et al. (2021) found that although subjective measures have been found to be relatively high in validity, their sensitivity to intrinsic load is maximised when used in combination with physiological measures. However, it was found that further research is needed on the effect of factors that increase extraneous load. This presents many challenges for collecting large amounts of data in authentic classroom contexts, given the intrusive nature of the equipment required to take many physiological measurements. One solution to this may lie in future CL research incorporating learner reflections or think-aloud protocols that provide learners the opportunity to elaborate on how they engaged with a particular task. As such, turning to a new source of data in understanding learner differences in CL may be a potentially fruitful line of inquiry for future research. The present study attempted to address this by including a brief learner training element to the instructions provided to students. However, the present study would have benefited from greater emphasis on this element to ensure greater validity of CL measures to support learners in differentiating between intrinsic and extraneous sources of load.

A third limitation of the present study relates to scratch paper use. Only the number of steps for which students relied on scratch paper was considered. The present data do not provide any insights into the qualitative differences in scratch paper use between testing modalities, nor whether these differences might be associated with more favourable test outcomes. Future research may also consider links between metacognitive strategies and scratch paper use in students of differing abilities. There are also exciting opportunities to explore the effect of learner training on the use of scratch paper and the way this translates to changes in cognitive strategies during learning and assessment tasks.

Finally, it is important to recognise that the measure of WMC may have provided a selective advantage to learners who are more confident in numeracy than those who are not. For instance, Carpenter and Alloway (2019) acknowledge the need for a wider array of WM measures to be taken to better understand the component of WM that are affected by different testing modes. Furthermore, Cowan et al. (2005) note that traditional storage-and-processing tasks like the reverse digit span make it difficult to determine whether a student's WMC score is determined by their attentional control, specific processing skills (i.e., verbal, linguistic, or spatial reasoning), or a combination of both. The justification for the inclusion of reverse digit span in the present study was based on (a) the numeracy component of the science content

being tested and (b) relative ease and efficiency of administration to allow data collection to occur within a single lesson. However, it is important to acknowledge that alternative measures of WM (such as N-back tasks, or visuospatial tasks such as the Corsi block recall) may provide more robust and insightful data about individual differences in WMC that are specifically affected by on-screen learning and assessment environments. Additionally, CLT posits that the experience of intrinsic load operates as a function of element interactivity and learner expertise. Thus, the extent to which a learner is placed in cognitive overload is dependent on not only their WMC but also domain knowledge. One important consideration for future analyses may be to include measures of learner expertise. In the context of school-aged students, this may include existing school grades for a given topic or subject.

Conclusion

To the knowledge of the authors, this is the first study to directly address test most effects on cognitive load in school-aged students whilst accounting for individual differences in working memory capacity. As the use of computers continues to make its way into the realm of assessment, it is important for educators to maintain a critical stance towards disruption-promising EdTech discourse (Selwyn, 2014) evoking “ideologically-charged common sense” (Friesen, 2008, p. 2). The present findings suggest that not all testing modalities affect all learners in equitable ways, especially when presenting school-aged learners with lower working memory capacities with more demanding tasks. It is also worth acknowledging that computer and multimedia environments can add extraneous element interactivity to a task (Sweller, 2020). In this study, both CBT and PBT versions were designed to appear equivalent to students in order to isolate testing mode effects. However, it is likely that computer-based learning and assessment environments will, in general, employ greater use of interactive elements and sources of digital distraction. Therefore, it is plausible that, in practice, the use of computer-based tasks adds a greater level of extraneous load than those identified here.

The present study highlights the need for research to consider optimal contextually appropriate cognitive load thresholds. This is consistent with previous research acknowledging the importance of alignment between costs of extraneous load induced by digital assessment environments and the potential benefits associated with the increases in germane processing these environments may invoke (Skulmowski & Xu, 2021). Additionally, these results highlight the importance of face validity of assessment in optimising student performance: although assessing students on a computer is more efficient from an administrative perspective, computer-based assessment environments are not universally beneficial. These findings are consistent with recent meta-analyses indicating that students achieve higher reading comprehension for information-based texts. Additionally, when under higher task demands induced by time constraints (i.e., extraneous load), the advantage of paper over computer-based texts increases (Clinton, 2019; Delgado et al., 2018). Ultimately, when dealing with younger learners, moving learning and assessment tasks from paper to computer is associated with a cost that is

likely to produce lower test scores, particularly as the intrinsic load of content increases and student WMC decreases. Because the primary goal of CLT is to generate instructional strategies that better facilitate learning (Sweller & Paas, 2017), the present research supports the interests of equitable assessment practices. This work underscores the importance of designing assessment practices to ensure teachers, educators, and policy makers are fairly and equitably testing content rather than creating assessments that place learners under an unnecessary testing load.

Author Contribution All authors contributed to the study conception and design. Material preparation, data collection, analysis, and draft preparation were performed by James Pengelley. All authors contributed to drafting of previous versions and approval of final version. Peter R. Whipp and Nina Rovis-Hermann are assigned to supervision.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions

Data Availability The data that support the findings of this study are not openly available due to reasons of sensitivity and are available from the corresponding author upon reasonable request.

Declarations

Competing Interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Ackerman, R., & Lauterman, T. (2012). Taking reading comprehension exams on screen or on paper? A metacognitive analysis of learning texts under time pressure. *Computers in Human Behavior*, 28(5), 1816–1828. <https://doi.org/10.1016/j.chb.2012.04.023>
- Anderson, J. R. (1983). *The architecture of cognition* (1st ed.). Psychology Press. <https://doi.org/10.4324/9781315799438>
- Anmarkrud, Ø., Andresen, A., & Bråten, I. (2019). Cognitive load and working memory in multimedia learning: Conceptual and measurement issues. *Educational Psychologist*, 54(2), 61–83. <https://doi.org/10.1080/00461520.2018.1554484>
- Annisette, L. E., & Lafreniere, K. D. (2017). Social media, texting, and personality: A test of the shal- lowing hypothesis. *Personality and Individual Differences*, 115, 154–158. <https://doi.org/10.1016/j.paid.2016.02.043>
- Ayres, P. (2001). Systematic mathematical errors and cognitive load. *Contemporary Educational Psychology*, 26(2), 227–248. <https://doi.org/10.1006/ceps.2000.1051>
- Ayres, P. (2006). Using subjective measures to detect variations of intrinsic cognitive load within prob- lems. *Learning and Instruction*, 16(5), 389–400. <https://doi.org/10.1016/j.learninstruc.2006.09.001>

- Ayres, P. (2015). State-of-the-Art research into multimedia learning: A commentary on Mayer's handbook of multimedia learning. *Applied Cognitive Psychology*, 29(4), 631–636. <https://doi.org/10.1002/acp.3142>
- Ayres, P., Lee, J. Y., Paas, F., & van Merriënboer, J. J. (2021). The validity of physiological measures to identify differences in intrinsic cognitive load. *Frontiers in Psychology*, 12, 702538. <https://doi.org/10.3389/fpsyg.2021.702538>
- Baddeley, A. D., & Hitch, G. J. (1994). Developments in the concept of working memory. *Neuropsychology*, 8(4), 485. <https://doi.org/10.1037/0894-4105.8.4.485>
- Batka, J. A., & Peterson, S. A. (2005). The effects of individual differences in working memory on multimedia learning. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 49(13), 1256–1260. <https://doi.org/10.1177/154193120504901309>
- Beckmann, J. (2010). Taming a beast of burden - on some issues with the conceptualisation and operationalisation of cognitive load. *Learning and Instruction*, 20(3), 250–264. <https://doi.org/10.1016/j.learninstruc.2009.02.024>
- Bennett, R. E., Braswell, J., Oranje, A., Sandene, B., Kaplan, B., & Yan, F. (2008). Does it matter if I take my mathematics test on computer? A second empirical study of mode effects in NAEP. *The Journal of Technology, Learning and Assessment*, 6(9).
- Bratfish, O., Borg, G., & Dornic, O. (1972). *Perceived item-difficulty in three tests of intellectual performance capacity*. Retrieved on October 6 2022 from <https://files.eric.ed.gov/fulltext/ED080552.pdf>
- Brünken, R., Steinbacher, S., Plass, J. L., & Leutner, D. (2002). Assessment of cognitive load in multimedia learning using dual-task methodology. *Experimental Psychology*, 49(2), 109–119. <https://doi.org/10.1023/B:TRUC.0000021812.96911.c5>
- Carpenter, R., & Alloway, T. (2019). Computer versus paper-based testing: Are they equivalent when it comes to working memory? *Journal of Psychoeducational Assessment*, 37(3), 382–394. <https://doi.org/10.1177/0734282918761496>
- Chen, O., Castro-Alonso, J. C., Paas, F., & Sweller, J. (2018). Extending cognitive load theory to incorporate working memory resource depletion: Evidence from the spacing effect. *Educational Psychology Review*, 30(2), 483–501. <https://doi.org/10.1007/s10648-017-9426-2>
- Choi, H.-H., Van Merriënboer, J. J., & Paas, F. (2014). Effects of the physical environment on cognitive load and learning: Towards a new model of cognitive load. *Educational Psychology Review*, 26, 225–244. <https://doi.org/10.1007/s10648-014-9262-6>
- Chu, H.-C. (2014). Potential negative effects of mobile learning on students' learning achievement and cognitive load—a format assessment perspective. *Journal of Educational Technology & Society*, 17(1), 332–344. <http://www.jstor.org/stable/jeductechsoci.17.1.332>. Accessed 6 Oct 2022
- Circuit Diagram Editor. (2022). Circuit Diagram. Retrieved October 9th 2021 from <https://www.circuit-diagram.org/editor/>
- Clinton, V. (2019). Reading from paper compared to screens: A systematic review and meta-analysis. *Journal of Research in Reading*, 42(2), 288–325. <https://doi.org/10.1111/1467-9817.12269>
- Collette, F., & Van der Linden, M. (2002). Brain imaging of the central executive component of working memory. *Neuroscience & Biobehavioral Reviews*, 26(2), 105–125. [https://doi.org/10.1016/S0149-7634\(01\)00063-X](https://doi.org/10.1016/S0149-7634(01)00063-X)
- Cowan, N., Elliott, E. M., Scott Saults, J., Morey, C. C., Mattox, S., Hismjatullina, A., & Conway, A. R. (2005). On the capacity of attention: Its estimation and its role in working memory and cognitive aptitudes. *Cognitive Psychology*, 51(1), 42–100. <https://doi.org/10.1016/j.cogpsych.2004.12.001>
- De Jong, T. (2010). Cognitive load theory, educational research, and instructional design: Some food for thought. *Instructional Science*, 38(2), 105–134. <https://doi.org/10.1007/s11251-009-9110-0>
- Debie, N., & Van De Leemput, C. (2014). What does germane load mean? An empirical contribution to the cognitive load theory. *Frontiers in Psychology*, 5, 1099. <https://doi.org/10.3389/fpsyg.2014.01099>
- Delgado, P., & Salmerón, L. (2021). The inattentive on-screen reading: reading medium affects attention and reading comprehension under time pressure. *Learning and Instruction*, 71, 101396. <https://doi.org/10.1016/j.learninstruc.2020.101396>
- Delgado, P., Vargas, C., Ackerman, R., & Salmerón, L. (2018). Don't throw away your printed books: A meta-analysis on the effects of reading media on reading comprehension. *Educational Research Review*, 25, 23–38. <https://doi.org/10.1016/j.edurev.2018.09.003>
- Ehrlick, S. P. (2014). Managing digital distraction: A pedagogical approach for dealing with wireless devices in the classroom. *Journal of Teaching and Education*, 3(3), 207–216.

- Endres, M. J., Houpt, J. W., Donkin, C., & Finn, P. R. (2015). Working memory capacity and redundant information processing efficiency. *Frontiers in Psychology*, 6. <https://doi.org/10.3389/fpsyg.2015.00594>
- Flanigan, A. E., & Titsworth, S. (2020). The impact of digital distraction on lecture note taking and student learning. *Instructional Science*, 48(5), 495–524. <https://doi.org/10.1007/s11251-020-09517-2>
- Friesen, N. (2008). Critical theory: ideology critique and the myths of E-learning. *Ubiquity*. 2008 (June). <https://doi.org/10.1145/1403922.1386860>
- Galy, E., Cariou, M., & Mélan, C. (2012). What is the relationship between mental workload factors and cognitive load types? *International Journal of Psychophysiology*, 83(3), 269–275. <https://doi.org/10.1016/j.ijpsycho.2011.09.023>
- Hardre, P. L., Crowson, H. M., Xie, K., & Ly, C. (2007). Testing differential effects of computer-based, web-based and paper-based administration of questionnaire research instruments. *British Journal of Educational Technology*, 38(1), 5–22. <https://doi.org/10.1111/j.1467-8535.2006.00591.x>
- Hatzigianni, M., Gregoriadis, A., & Fleer, M. (2016). Computer use at schools and associations with social-emotional outcomes—a holistic approach. Findings from the longitudinal study of Australian Children. *Computers & Education*, 95, 134–150.
- Jiang, D., & Kalyuga, S. (2020). Confirmatory factor analysis of cognitive load ratings supports a two-factor model. *Tutorials in Quantitative Methods for Psychology*, 16(3), 216–225. <https://doi.org/10.20982/tqmp.16.3.p216>
- Kalyuga, S. (2011). Cognitive load theory: How many types of load does it really need? *Educational Psychology Review*, 23(1), 1–19. <https://doi.org/10.1007/s10648-010-9150-7>
- Kalyuga, S., & Singh, A.-M. (2016). Rethinking the boundaries of cognitive load theory in complex learning. *Educational Psychology Review*, 28(4), 831–852. <https://doi.org/10.1007/s10648-015-9352-0>
- Kessels, R. P., van Den Berg, E., Ruis, C., & Brands, A. M. (2008). The backward span of the corsi block-tapping task and its association with the WAIS-III digit span. *Assessment*, 15(4), 426–434. <https://doi.org/10.1177/1073191108315611>
- Klepsch, M., Schmitz, F., & Seufert, T. (2017). Development and validation of two instruments measuring intrinsic, extraneous, and germane cognitive load. *Frontiers in Psychology*, 8. <https://doi.org/10.3389/fpsyg.2017.01997>
- Korbach, A., Brünken, R., & Park, B. (2018). Differentiating different types of cognitive load: A comparison of different measures. *Educational Psychology Review*, 30(2), 503–529. <https://doi.org/10.1007/s10648-017-9404-8>
- Latini, N., Bräten, I., & Salmerón, L. (2020). Does reading medium affect processing and integration of textual and pictorial information? A multimedia eye-tracking study. *Contemporary Educational Psychology*, 62, 101870. <https://doi.org/10.1016/j.cedpsych.2020.101870>
- Likourezos, V., Kalyuga, S., & Sweller, J. (2019). The variability effect: When instructional variability is advantageous. *Educational Psychology Review*, 31(2), 479–497. <https://doi.org/10.1007/s10648-019-09462-8>
- Logan, T. (2015). The influence of test mode and visuospatial ability on mathematics assessment performance. *Mathematics Education Research Journal*, 27(4), 423–441. <https://doi.org/10.1007/s13394-015-0143-1>
- Maerlender, A. C., Wallis, D. J., & Isquith, P. K. (2004). Psychometric and behavioral measures of central auditory function: The relationship between dichotic listening and digit span tasks. *Child Neuropsychology*, 10(4), 318–327. <https://doi.org/10.1080/09297040490909314>
- Martin, S. (2014). Measuring cognitive load and cognition: Metrics for technology-enhanced learning. *Educational Research and Evaluation*, 20(7–8), 592–621. <https://doi.org/10.1080/13803611.2014.997140>
- Mayer, R. E. (2003). The promise of multimedia learning: Using the same instructional design methods across different media. *Learning and Instruction*, 13(2), 125–139. [https://doi.org/10.1016/S0959-4752\(02\)00016-6](https://doi.org/10.1016/S0959-4752(02)00016-6)
- Mayer, R. E. & Fiorella, L. (2021). Principles for managing essential processing in multimedia learning: segmenting, pretraining, and modality principles. In Mayer, R. E., & Fiorella, L. (Eds.), *The Cambridge handbook of multimedia learning* (pp. 243–260). Cambridge University Press. <https://doi.org/10.1017/9781108894333>
- Mayes, D., Sims, V., & Koonce, J. (2001). Comprehension and workload differences for VDT and paper-based reading. *International Journal of Industrial Ergonomics*, 28(6), 367–378. [https://doi.org/10.1016/S0169-8141\(01\)00043-9](https://doi.org/10.1016/S0169-8141(01)00043-9)

- Mueller, P. A., & Oppenheimer, D. M. (2014). The pen is mightier than the keyboard: Advantages of longhand over laptop note taking. *Psychological Science*, 25(6), 1159–1168. <https://doi.org/10.1177/0956797614524581>
- National Assessment Programme. (2016). FAQs, NAPLAN - General. ACARA. Retrieved July 16 from <https://www.nap.edu.au/naplan/faqs/naplan-general>
- Naismith, L. M., Cheung, J. J., Ringsted, C., & Cavalcanti, R. B. (2015). Limitations of subjective cognitive load measures in simulation-based procedural training. *Medical Education*, 49(8), 805–814. <https://doi.org/10.1111/medu.12732>
- Noyes, J., & Garland, K. (2003). VDT versus paper-based text: Reply to Mayes, Sims and Koonce. *International Journal of Industrial Ergonomics*, 31(6), 411–423. [https://doi.org/10.1016/S0169-8141\(03\)00027-1](https://doi.org/10.1016/S0169-8141(03)00027-1)
- Noyes, J., Garland, K., & Robbins, L. (2004). Paper-based versus computer-based assessment: Is workload another test mode effect? *British Journal of Educational Technology*, 35(1), 111–113. <https://doi.org/10.1111/j.1467-8535.2004.00373.x>
- OECD, (2010). PISA Computer-Based Assessment of Student Skills in Science, PISA, OECD Publishing, Paris. <https://doi.org/10.1787/9789264082038-en>
- Ophir, E., Nass, C., & Wagner, A. D. (2009). Cognitive control in media multitaskers. *Proceedings of the National Academy of Sciences*, 106(37), 15583–15587. <https://doi.org/10.1073/pnas.0903620106>
- Paas, F., & Van Merriënboer, J. (1993). The efficiency of instructional conditions: An approach to combine mental effort and performance measures. *Human Factors*, 35(4), 737–743. <https://doi.org/10.1177/001872089303500412>
- Paas, F., Van Merriënboer, J., & Adam, J. (1994). Measurement of cognitive load in instructional research. *Perceptual and Motor Skills*, 79(1), 419–430. <https://doi.org/10.2466/pms.1994.79.1.419>
- Pajares, F. (2004). Gender differences in mathematics self-efficacy beliefs. In A. Gallagher & J. Kaufman (Eds.), *Gender Differences in Mathematics: An Integrative Psychological Approach* (pp. 294–315). Cambridge University Press. <https://doi.org/10.1017/CBO9780511614446.015>
- Park, B., Moreno, R., Seufert, T., & Brünken, R. (2011). Does cognitive load moderate the seductive details effect? A multimedia study. *Computers in Human Behavior*, 27(1), 5–10. <https://doi.org/10.1016/j.chb.2010.05.006>
- Park, B., Korbach, A., & Brünken, R. (2015). Do learner characteristics moderate the seductive-details-effect? A cognitive-load-study using eye-tracking. *Journal of Educational Technology & Society*, 18(4), 24–36.
- Prisacari, A., & Danielson, J. (2017). Computer-based versus paper-based testing: Investigating testing mode with cognitive load and scratch paper use. *Computers in Human Behavior*, 77, 1–10. <https://doi.org/10.1016/j.chb.2017.07.044Get>
- Qualtrics. (2020). *Qualtrics XM*. (Version October, 2021) Qualtrics. <https://www.qualtrics.com>
- Raaijmakers, S., Baars, M., Schaap, L., Paas, F., & Van Gog, T. (2017). Effects of performance feedback valence on perceptions of invested mental effort. *Learning and Instruction*, 51, 36–46. <https://doi.org/10.1016/j.learninstruc.2016.12.002>
- Ronconi, A., Veronesi, V., Mason, L., Manzione, L., Florit, E., Anmarkrud, Ø., & Bråten, I. (2022). Effects of reading medium on the processing, comprehension, and calibration of adolescent readers. *Computers and Education*, 185, 104520. <https://doi.org/10.1016/j.compedu.2022.104520>
- School Curriculum and Standards Authority. (2014). *Naplan background*. Retrieved October 8 2022 from <https://k10outline.scsa.wa.edu.au/home/assessment/testing/naplan>
- Schmeck, A., Opfermann, M., Van Gog, T., Paas, F., & Leutner, D. (2015). Measuring cognitive load with subjective rating scales during problem solving: Differences between immediate and delayed ratings. *Instructional Science*, 43(1), 93–114. <https://doi.org/10.1007/s11251-014-9328-3>
- Schnotz, W. (2010). Reanalyzing the expertise reversal effect. *Instructional Science*, 38(3), 315–323.
- Schnotz, W., & Kürschner, C. (2007). A reconsideration of cognitive load theory. *Educational Psychology Review*, 19(4), 469–508. <https://doi.org/10.1007/s10648-007-9053-4>
- Selwyn, N. (2014). Distrusting educational technology: Critical questions for changing times. *Routledge, Taylor & Francis Group*. <https://doi.org/10.4324/9781315886350>
- Selwyn, N. (2016). Minding our language: why education and technology is full of bullshit... and what might be done about it. *Learning, Media and Technology*, 41(3), 437–443. <https://doi.org/10.1080/17439884.2015.1012523>
- Sidi, Y., Shpigelman, M., Zalmanov, H., & Ackerman, R. (2017). Understanding metacognitive inferiority on screen by exposing cues for depth of processing. *Learning and Instruction*, 51, 61–73. <https://doi.org/10.1016/j.learninstruc.2017.01.002>

- Skulmowski, A., & Xu, M. (2021). Understanding cognitive load in digital and online learning: A new perspective on extraneous cognitive load. *Educational Psychology Review*, 33(2), 171–196. <https://doi.org/10.1007/s10648-021-09624-7>
- Solhjo, S., Haigney, M. C., McBee, E., van Merriënboer, J. J., Schuwirth, L., Artino, A. R., Battista, A., Ratcliffe, T. A., Lee, H. D., & Durning, S. J. (2019). Heart rate and heart rate variability correlate with clinical reasoning performance and self-reported measures of cognitive load. *Scientific Reports*, 9(1), 1–9. <https://doi.org/10.1038/s41598-019-50280-3>
- St Clair-Thompson, H. L., & Gathercole, S. E. (2006). Executive functions and achievements in school: Shifting, updating, inhibition, and working memory. *The Quarterly Journal of Experimental Psychology*, 59(4), 745–759. <https://doi.org/10.1080/17470210500162854>
- Sweller, J. (2020). Cognitive load theory and educational technology. *Educational Technology Research and Development*, 68(1), 1–16. <https://doi.org/10.1007/s11423-019-09701-3>
- Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4
- Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational Psychology Review*, 22(2), 123–138. <https://doi.org/10.1007/s10648-010-9128-5>
- Sweller, J. (2018). Measuring cognitive load. *Perspectives on Medical Education*, 7(1), 1–2. <https://doi.org/10.1007/s40037-017-0395-4>
- Sweller, J., Ayres, P., & Kalyuga, S. (2011). Cognitive load theory. *Springer*. <https://doi.org/10.1007/978-1-4419-8126-4>
- Sweller, J., & Chandler, P. (1994). Why some material is difficult to learn. *Cognition and Instruction*, 12(3), 185–233. https://doi.org/10.1207/s1532690xci1203_1
- Sweller, J., van Merriënboer, J. J., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review*, 31, 1–32. <https://doi.org/10.1007/s10648-019-09465-5>
- Sweller, J., & Paas, F. (2017). Should self-regulated learning be integrated with cognitive load theory? A commentary. *Learning and Instruction*, 51, 85–89. <https://doi.org/10.1016/j.learninstruc.2017.05.005>
- Sweller, J., Van Merriënboer, J. J., & Paas, F. G. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, 10(3), 251–296. <https://doi.org/10.1023/A:1022193728205>
- Szulewski, A., Gegenfurtner, A., Howes, D. W., Sivilotti, M. L., & van Merriënboer, J. J. (2017). Measuring physician cognitive load: Validity evidence for a physiologic and a psychometric tool. *Advances in Health Sciences Education*, 22(4), 951–968. <https://doi.org/10.1007/s10459-016-9725-2>
- The jamovi project (2022). *jamovi* (Version 2.2.5) [Computer Software]. Retrieved from <https://www.jamovi.org>
- Van Gog, T., Kirschner, F., Kester, L., & Paas, F. (2012). Timing and frequency of mental effort measurement: Evidence in favour of repeated measures. *Applied Cognitive Psychology*, 26(6), 833–839. <https://doi.org/10.1002/acp.2883>
- Van Merriënboer, J. J., Kester, L., & Paas, F. (2006). Teaching complex rather than simple tasks: Balancing intrinsic and germane load to enhance transfer of learning. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 20(3), 343–352. <https://doi.org/10.1002/acp.1250>
- Vassallo, S., & Warren, D. (2017). Use of technology in the classroom. In D. Warren & G. Daraganova (Eds.), *Growing Up In Australia – The Longitudinal Study of Australian Children, Annual Statistical Report 2017* (pp. 99–112). Australian Institute of Family Studies.
- Whelan, R. R. (2007). Neuroimaging of cognitive load in instructional multimedia. *Educational Research Review*, 2(1), 1–12. <https://doi.org/10.1016/j.edurev.2006.11.001>
- Wilde, N. J., Strauss, E., & Tulskey, D. S. (2004). Memory span on the Wechsler Scales. *Journal of Clinical and Experimental Neuropsychology*, 26(4), 539–549. <https://doi.org/10.1080/13803390490496605>