



Designing computer-based tests: design guidelines from multimedia learning studied with eye tracking

K. J. H. Dirkx¹ · I. Skuballa^{1,2} · C. S. Manastirean-Zijlstra² · H. Jarodzka¹

Received: 12 November 2018 / Accepted: 10 May 2021 / Published online: 2 September 2021
© The Author(s) 2021

Abstract

The use of computer-based tests (CBTs), for both formative and summative purposes, has greatly increased over the past years. One major advantage of CBTs is the easy integration of multimedia. It is unclear, though, how to design such CBT environments with multimedia. The purpose of the current study was to examine whether guidelines for designing multimedia *instruction* based on the Cognitive Load Theory (CLT) and Cognitive Theory of Multimedia Learning (CTML) will yield similar effects in CBT. In a within-subject design, thirty-three vocational students completed a computer-based arithmetic exam, in which half of the items were presented in an original design format, and the other half was redesigned based on the CTML principles for instructional design (i.e., adapted format). Results showed that applying CTML principles to a CBT decreased the difficulty of the test items, i.e., students scored significantly higher on the adapted test items. Moreover, eye-tracking data showed that the adapted items required less visual search and increased attention for the question and answer. Finally, cognitive load, measured as silent pauses during a secondary think-aloud task, decreased. Mean fixation duration (a different indicator of cognitive load), however, did not significantly differ between adapted and original items. These results indicate that applying multimedia principles to CBTs can be beneficial. It seems to prevent cognitive overload and helps students to focus on important parts of the test items (e.g., the question), leading to better test results.

Keywords Computer-based testing · Eye tracking · Mental effort · Multimedia learning · Design

Introduction

Computers are frequently used in education to support and assess students' knowledge. The increasing use of computers for both formative and summative testing has many advantages. Computers provide for example an excellent medium to include different

✉ H. Jarodzka
Halszka.Jarodzka@ou.nl

¹ Faculty of Educational Sciences Heerlen, Open Universiteit Nederland, NL - P.O. Box 2960, 6401 DL Heerlen, The Netherlands

² ROC Flevoland, MBO College Poort, Almere, The Netherlands

representations (e.g., videos, text, and pictures) and response formats (e.g., drag-and-drop questions, interactive questions) in tests (see Basu et al., 2007 for an overview; Mayer, 2014a, 2019). This makes computer-based tests (CBT) more engaging, effective, and entertaining as compared to paper-based tests (Azabdaftari & Mozaheb, 2012; Başoğlu & Akdemir, 2010; Basu et al., 2007; Chua & Don, 2013; Kaplan-Rakowski, & Loranc-Paszylk, 2017) and increases student motivation (Lin & Yu, 2017).

Despite these benefits, concrete design guidelines on how best to present multimedia content in test items are lacking, thus far. Research on *learning* with multimedia has, however, shown that the concrete design of *learning materials* (e.g., when to insert which type of pictures) has a strong influence on how learners process the presented information and thus, on the learning outcome itself (Alemdag & Cagiltay, 2018). Although learning is not the same as testing (i.e., the former requires information intake and the latter information retrieval) both require sensory and cognitive processing of information that is subjected to the same characteristics and restrictions of the human information-processing-system (Baddeley, 1992). Even more, to understand what is given and what is asked in a testing item, information intake is the very first step. In this light, Kirschner et al. (2016) suggested that testing involves similar cognitive processes as learning (i.e., schema creation and retrieval) and that thus, the design of the test items will play a vital role for its processing (cf. Jarodzka et al., 2015; Ögren et al., 2016). Consequently, Kirschner et al. (2016) suggested investigating and determining to which extent cognitive theories on *learning* with multimedia—the Cognitive Theory of Multimedia Learning (CTML; Mayer, 2014a, b) and the Cognitive Load Theory (CLT; Sweller, 2011; Sweller et al., 1998) – apply to testing.

Human cognitive architecture and information processing

Designers' understanding of how humans learn is reflected in their composition of multimedia messages and instructional materials. One crucial maxim is that the cognitive system of the learner should not be overloaded and meaningful cognitive processing should be supported through instructional design (Mayer, 2014a b; Paas & Sweller, 2014). Both theories, CLT (Sweller, 2011; Sweller et al., 1998) and CTML (Mayer, 2009, 2014a), make specific assumptions about how people process multimedia material.

The CTML perspective

Against the background of cognitive science, CTML builds on three pillars describing human learning (Mayer, 2009). First, incoming visual and auditory information is processed through separate channels (dual-channel assumption; Baddeley, 1992; Paivio, 1986). Secondly, the human cognitive system is limited in its amount of information that can be processed simultaneously (limited capacity assumption; Baddeley, 1992; Chandler & Sweller, 1991). Finally, the learner must engage in active processing to acquire knowledge (Mayer, 2009). Active processing refers to drawing attention to information, selection of relevant information from the presentations, organization of selected information into coherent mental models, and the integration of the mental models with prior knowledge from long-term memory (Mayer, 2009, 2014a).

There are three demands on the learner's cognitive capacity that can be ascribed to external sources or sources inert to the learning task (Mayer, 2009, 2014a): extraneous processing is not related to the instructional objective but rather the processing of task-irrelevant extraneous material, essential processing is the mental representation of the essential

material as presented to the learner, and generative processing is the process of making sense of the material. Extraneous and essential processing can result in overload when they exceed the learner's cognitive capacities. The cognitive activities required for meaningful learning, i.e., selecting, organizing, and integrating, are linked to those load kinds. Selection is associated with essential load; organization and integration are associated with generative load (Mayer, 2009). Therefore, good instructional design should support these processes and hence meaningful learning through (a) the reduction of extraneous processing due to insufficient presentation, (b) management of essential processing, and (c) fostering of generative processing. Each of these areas is defined by specific design principles. The present study focuses on the reduction of extraneous processing.

The CLT perspective

Similarly, the CLT explains load related to the learning tasks which affect the learner and offers a theoretical framework of the human cognitive architecture from which design principles can be derived (Sweller et al., 2011). The human cognitive structure consists of a long-term memory (LTM) and a working memory (WM). The capacity of the long-term memory is unlimited, whereas the working memory is limited in its capacity and duration (Choi et al., 2014; Sweller et al., 1998). Note that similar assumptions are also made within the CTML (Mayer, 2009). In particular new and unfamiliar information challenge the limited capacity of the working memory (Choi et al., 2014). These challenges can be overcome when the new information can be easily connected and integrated with pre-existing information from long-term memory. The broader the prior knowledge base of a learner, the easier she will manage new incoming information. Cognitive schemas are an efficient way to organize and store knowledge. The construction of cognitive schemas refers to the chunking of similar information elements into one single element which is stored in LTM. The retrieval of one element from LTM is less load-intensive and simultaneously activates other information pieces within the schema which can be linked with the new information piece in working memory. Through practice, the learner can reach a level of automation where the schema can be used unconsciously thus successfully tackling the limitations of WM (Choi et al., 2014).

Within CLT, two load types can be distinguished: intrinsic load and extrinsic load (Choi et al., 2014; Sweller et al., 2019). Intrinsic load refers to the given and unchangeable complexity of the learning task, specifically the amount of interacting information elements that must be processed to master the learning task (Choi et al., 2014; Sweller et al., 2019). Extraneous load, on the other hand, is determined by the way the information is presented and the instructions provided to the learner. Unlike intrinsic load, the extraneous load can be changed through instructional design. Previous models have also introduced germane load as a third distinct load type defined as the cognitive load required to learn (Sweller et al., 2019). However, revised versions of the CLT consider intrinsic and germane cognitive load "closely intertwined" (Kalyuga, 2011; Sweller et al., 2019, p. 264), and rather call the WM resources dealing with intrinsic load "germane resources" (Choi et al., 2014, p. 227). The load kinds in CTML can be understood within the CLT as follows: essential cognitive processing corresponds to intrinsic cognitive load, generative cognitive processing corresponds to germane cognitive load, and finally extraneous processing to extraneous load (Mayer, 2009).

The measurement of cognitive load can be conceptualized through mental load, mental effort, and performance. Task performance is considered an indirect indicator of the

cognitive load because the amount of cognitive load during task processing is interpreted post-hoc in relation to the measured task performance (Korbach et al., 2017b). Measuring cognitive load through a secondary task is considered an objective and reliable method (Sweller et al., 2011; Korbach et al., 2017b). It obtains behavioral data from a secondary task that is performed by the learner in addition to a primary task (dual-task methodology). The secondary task represents experienced cognitive load during a task, whereas the subjective measures indicate load after the completion of the task (Sweller et al., 2011). Previous work has demonstrated that speech analysis can serve as a promising cognitive load measurement (Yin & Chen, 2007). Speech-based features such as silent pauses are considered potential indicators of the level of cognitive load: participants make more pauses when they experience higher load. In addition, pauses in speech are associated with fixation durations which can be considered another indicator of cognitive load (cf. Jarodzka et al., 2015). Despite many advantages, the objective measures are less widely used because the subjective measures can be easier and quicker administered.

Multimedia in testing

Recently, there have been several studies aiming to reduce the extraneous load in testing by applying multimedia principles to computer-based testing (e.g., Jarodzka et al., 2015; Lindner et al., 2017; Lindner et al., 2016). When solving test questions, testees need to read test item stimuli and item stems, interpret visuals, and type in the correct response. It is assumed that CTLM and CLT can inform test design for computer-based test items (Beddow, 2018; Kirschner et al., 2016) but recent studies show applying CTML principles to testing does not always yield the expected results (e.g., Ögren et al., 2016). One of the reasons explicated in the literature is that the objectives of the constructed materials in instructional design and assessment are different (Beddow, 2018). Instructional materials are designed to produce instructional outcomes (i.e., acquisition and retention of knowledge) whereas assessment materials are designed to assess them.

A design principle that gained most attention thus far in the research focusing on testing is the basic principle of combining text and pictures (i.e., ‘multimedia effect’;). The empirical evidence confirms that there is a positive effect of adding pictures to test items on performance (see Butcher, 2014 for an overview; Lindner et al., 2017a; Lindner et al., 2016) which is explained as follows: Pictures seem to serve as mental scaffolds that support comprehension and decision making behavior during testing (Lindner et al., 2017a).

However, research also shows that pictures are not always helpful in CBT (Anglin et al., 2004; Jarodzka et al., 2015; Ögren et al., 2016). For instance, in a study by Jarodzka et al. (2015) on standardized testing material, pictures were either presented away from the text (as the testing material was originally designed) or positioned in close proximity to the text (contiguity principle). Analyses of eye-tracking data showed that the optimized, integrated presentation format guided students’ attention to inspecting all of the multimedia material. This, however, did not improve testing performance. Post-hoc analyses of the content of the standard testing material revealed that a substantial amount of the provided information (pictorial and textual) in the question was not relevant for solving the test item and thus violated the coherence principle (Mayer, 2014b).

A very likely explanation for these findings can again be found in the learning literature. Here, irrelevant information is known to hamper performance (cf. negative effect of seductive details on learning: Abercrombie, 2013; or coherence principle: Mayer, 2014b). An interesting finding of Jarodzka et al. (2015) though was that a longer visual inspection of

the question and answer itself (as opposed to the picture and explanatory text), was positively related to better test performance. Likewise, in the study by Ögren et al. (2016) on vector calculus with and without graphs, no overall multimedia effect was found. In that study, students were presented items containing a statement about a presented formula. In half of the items, the formula was presented in the test question on the right side of the screen and in the other half of the items, the formula was accompanied by a relevant visual representation of that formula on the right side of the screen. The results showed that students looked proportionally less at the test question containing the written formula when a graph was present and students experienced more cognitive load in the multimedia condition, as indicated by more silent pauses in thinking aloud. Further analysis revealed, however, again that the more students looked at the question, the better they performed. Thus, it seems that pictures and graphs draw students' attention and cognitive capacities away from focusing on the question and hinder, under certain conditions, performance. This could mean that applying *one* of the multimedia principles in isolation to reduce the extraneous load in testing (e.g., contiguity) is too simplistic and provides a distorted picture. Instead, it seems more valid to suppose that when adding pictures to test questions, the pictures need to be relevant (i.e., multimedia principle and seductive details), pictures and text should be well integrated (i.e., contiguity principle), and redundant information should be resolved (i.e., redundancy principle). However, no research thus far has combined several multimedia principles to optimize *authentic* computer-based test items. But, to be able to investigate whether the multimedia principles are worthwhile for multimedia assessment, this is an important investigation to conduct as a starting point for more in-depth research and theory building that can inform practitioners.

Visual processing of multimedia material

Prior research (e.g., Ögren et al., 2016) has shown that eye-tracking is a useful technique to gain a better understanding of the effects of design modifications (see Alemdag & Cagiltay, 2018 for a review; Lin et al., 2017) as it provides insights into the cognitive and perceptual processes that the design evokes in the testee. Eye-tracking enables researchers to measure processes underlying learning or test performance in a more objective and online manner than, for example, self-reporting (Van Gog & Jarodzka, 2013). Eye-tracking captures two types of measurements: fixations and saccades. Fixations reflect where the learner is attending to whereas saccades show the change in focus of visual attention (cf. Holmqvist et al., 2011). The longer a person is fixating (operationalized as e.g., total fixation duration) on an area of interest, the more processing effort this area evokes (Underwood et al., 2004). Besides fixation durations, the number of revisits to certain on-screen areas can also provide useful information as it can indicate the extent to which the learner or testee engages in integration processes of the different elements (Alemdag & Cagiltay, 2018).

Another measure that is more and more used to measure cognitive load is dual-task performance (cf. Jarodzka et al., 2015; Yin & Chen, 2007). Under dual-task conditions, performance on the secondary task (i.e., thinking aloud) or both the primary (e.g., successful completion of the computer-based test questions) and the secondary task will suffer — for example, indicated by increased silent pauses (Yin & Chen, 2007)—when the task itself imposes high levels of cognitive load. Used as such, dual-task performance enables researchers to measure cognitive load in an objective and online way (e.g., Brünken et al., 2003; Ericsson & Simon, 1993; Korbach et al., 2017; van Gog et al., 2009).

Present study

The present study focuses on the extraneous load type and the principles aiming to reduce extraneous processing by omitting extraneous materials in the task. Learners experience extraneous processing overload when extraneous materials are irrelevant to the learning task but still attract their attention so that they are distracted from essential and generative processing (Mayer, 2009). Coherence, signaling, and spatial and temporal contiguity principles are examples of evidence-informed principles that were shown effective in reducing extraneous processing when designing learning materials (Mayer, 2009, 2014b). However, similar principles should also be key in designing testing materials (Beddow, 2018). The present study aims to investigate whether design principles that reduce the extraneous load in learning can be transferred to testing situations and will reveal similar positive effects.

The effect of an extraneous load reduction in *testing* is expected to be similar to the one in *learning* with multimedia: Since extraneous processing is linked to the presentation format, we assume that presenting an assessment task in a more beneficial way by applying the multimedia principles should reduce cognitive load bound by extrinsic load due to extraneous material. Reducing extrinsic load should free capacities in working memory for essential processing and release cognitive resources for retrieval of prior knowledge from LTM. Given these considerations, reducing extraneous processing in testing situations should result in more correct responses. Based on existing literature we formulated three hypotheses (cf. Lindner et al., 2017):

Hypothesis 1: Performance hypothesis Adapted items will result in lower item difficulty (i.e., higher item solving probability) than original items.

Hypothesis 2: Visual search hypothesis Adapted items lead to longer relative viewing and more revisits to the question and answer (referred to in this paper as item stem) and shorter relative viewing times and less revisits to the explanatory text and/or picture (referred to in this paper as item stimulus) than original items.

Hypothesis 3: Cognitive load hypothesis Cognitive load is lower for the adapted items as compared to the original items. Cognitive load is operationalized as mean fixation duration and duration of silent pauses.

Methods and materials

Participants and design

The sample consisted of thirty-three (44% female) Dutch vocational education students. Four students were excluded from the analyses because of inaccuracy in the eye-tracking data, resulting in valid data of 29 students. Power analysis with G-power 3.1.9.4 showed that this should be a sufficiently large sample size to detect an effect. The students were in their first or second year of schooling and were between 16 and 28¹ years old

¹ 80% of the students were between 16 and 21 years old. Despite the large standard deviation, age differences did not lead to outliers in test scores or cognitive load.

($M = 18.77$ years; $SD = 3.05$). Age ranges were high because in vocational training education students can drop-out from schooling when they are 18 but return to school when they are older. All students followed math classes at the minimal level (i.e., 2F) required to manage in society as defined by the Dutch national governance (Rijksoverheid, 2018). Participation in the study was voluntary and students were rewarded with a small treat afterwards. For the present study a within-subject experiment—with test item design as an independent factor—was conducted. Dependent variables were item difficulty, visual search (fixation duration and revisits), and cognitive load (mean fixation duration and silent pauses).

Testing material

The test used in the present study was an adaptation of an authentic *standardized* computer-based mathematics test used nationwide in Dutch vocational education. The items were presented in a self-paced way to students. The test used here contained 10 items and was presented to the students in the digital assessment environment FACET 4.0. There were two versions of the test. In version 1, the even items were in an adapted format and the uneven items were in original format whereas, in version 2, the even items retained its original format and uneven items were adapted. Participants were randomly assigned to one of the two test versions. It is important to note that for the current study, we optimized the original items according to those multimedia principles that aim at reducing extraneous cognitive load. This means that — depending on how the original, nationwide used, items were constructed—several modifications were implemented. For post-hoc analyses, we made three clusters of items. For the first cluster of items ($n = 4$), no elements were deleted but elements that belonged together were presented closely near each other and in a contingent order (i.e., contingency principle). For the second cluster ($n = 3$) some seductive elements were replaced by text and if the item contained redundant information, that information was removed. After adaptation, the items still contained a visual element that contained necessary information. In addition, information was presented in a more contingent manner on one side of the screen only (i.e., contingency, redundancy, and coherence principle). For the third cluster of items ($n = 3$), all seductive visual elements and redundant were removed and again information was presented in a more contingent manner on one side of the screen only (i.e., contingency, redundancy, and coherence). Figure 1 shows an example item of each of the three clusters including the original and adapted item.


Measures

Item-difficulty

The computer-based testing environment, FACET 4.0, provided individual performance scores per test item from which we were able to calculate item difficulty (total score on the item/number of participants).


Eye-tracking parameters

Relevant eye-tracking parameters were total and mean fixation durations and revisits on relevant on-screen elements. Therefore, these parameters were assigned to certain areas of interest (AOIs). The names of the AOIs were based on Beddow et al. (2009) and




Het is 24 juli en je kijkt rond 9 uur 's ochtends hoe lang het nog duurt voordat de Olympische Spelen beginnen.

Original



Het is 24 juli en je kijkt rond 9 uur 's ochtends hoe lang het nog duurt voordat de Olympische Spelen beginnen.


Adapted



Totale hoogte: 259 cm.

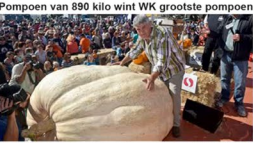
Bij een brug staat dat de maximale doorhoogte 3,0 meter is.

Original



Bij een brug staat dat de maximale doorhoogte 3,0 meter is.


Adapted



Pompoen van 890 kilo wint WK grootste pompoen.

Deze pompoen van 890 kg wint het WK grootste pompoen. De pompoen die vorig jaar won, woog 1055 kg.

Original



Een pompoen van 890 kg wint het WK grootste pompoen. De pompoen die vorig jaar won, woog 1055 kg.

Adapted

Fig. 1 Examples of item manipulations

Explanatory text

Je bestelt 15 foto's van het formaat 10x15.

Question

Hoeveel moet je foto's betalen?

€ **Answer**

foto's afdrukken

prijsoverzicht (in euro)

formaat (in cm)	tot 49 stuks	v.a. 50 stuks	v.a. 75 stuks	v.a. 120 stuks	v.a. 250 stuks	v.a. 500 stuks
10x10	0,20	0,19	0,17	0,13	0,11	0,09
10x15	0,20	0,19	0,17	0,13	0,11	0,09
11,4x15	0,23	0,21	0,19	0,14	0,12	0,10
13x13	0,45					
12,7x19	0,45					
15x20	0,55					
20,3x30	1,60					

Visualisation

Item

Fig. 2 Defined AOIs in one of the items as an example (explanatory text and visualization are combined in 'item stimulus' and question and answer are combined in 'item stem')

were: 'item stimulus' (i.e., explanatory text+visualization) and 'item stem' (i.e., question+answer). In Fig. 2 the AOI's for one item are presented as an example.

Cognitive load

There were two measures for cognitive load. Mean fixation duration was calculated as a first measure for cognitive load. Mean fixation duration is the mean duration of all fixation durations on a certain area of interest during a trial. Higher cognitive load has been shown to be related to increased fixation durations in prior research (Jarodzka et al., 2015; cf. Hyönä, 2010; cf., van Gog et al., 2009). This parameter was measured per item.

Second, silent pauses were used as a speech-based measurement of cognitive load (Yin & Chen, 2007). Participants were trained and instructed in thinking aloud according to Ericsson and Simon (1993) while completing the test questions. They were asked to "*verbalize everything that comes to mind, and disregard the experimenter's presence in doing so*" (Ericsson & Simon, 1993; van Gog et al., 2005). If they were silent for 15 s, they were reminded to keep thinking aloud (van Gog et al., 2005). With the software Audacity, duration of silent pauses longer than two seconds were identified and registered in Excel (Jarodzka et al., 2015). We counted the frequency of pauses lasting more than two seconds (cf. Ericsson & Simon, 1993).

Apparatus

Eye movements were recorded using a remote, video-based eye-tracking system (SMI RED; 250 Hz sampling rate). The apparatus was placed in a quiet room in school and students were tested individually. Participants sat at approximately 60 cm away from the screen.

Before data collection, the system was calibrated using a 5-point pulsing calibration image and subsequent validation. The computer-based test items were presented on a standard monitor of 22-inch, with a 1680×1050 picture resolution using the screen recording function of the software Experiment Center 3.7 from SensoMotoric Instruments (SMI; Teltow, Germany). Mean calibration accuracy was $M_x=0.56$ ($SD=0.19$) degrees and $M_y=0.52$ ($SD=0.34$) degrees of visual angle. Average tracking ratio was 81.1 ($SD=9.29$). The data were analyzed with BeGaze version 3.7, from SensoMotoric Instruments (SMI; Teltow, Germany).

The silent pauses longer were recorded using the microphone of an external Logitech Pro 9000 Business webcam. The camera itself was covered so no video material of the participant was collected.

Procedure

Students were tested in a quiet room at their school in individual sessions. After signing informed consent, students were instructed to think aloud using an example item (i.e., "*please think-out loud while solving 20×11?*", Ericsson & Simon, 1993). Next, the eye-tracking equipment was calibrated, and after successful calibration students were once more reminded to think aloud and the CBT began. Students worked at their own pace but had a maximum of 30 min to complete the test. Students were not allowed to use their calculators. During test completion, students were reminded to keep on thinking aloud when

pauses of 15 s occurred. After the test was completed, students were rewarded with a small treat (i.e., chocolate bar).

Data analyses

The test-scores were transferred to SPSS version 24 and in accordance with classical testing theory, we calculated mean item difficulty as the proportion (p) of participants who got the item correct. A one-way repeated measures ANOVA with ‘format’ as within-subject variable and ‘difficulty’ as the dependent variable was used to analyze the results. For the AOIs, relative fixation duration and revisits were calculated as indicators for visual search. Relative fixation duration was calculated by total fixation duration on AOI/total fixation time on the item. We used MANOVA to analyze shifts in attention allocation between adapted and original items on the different AOIs. A one-way repeated measures ANOVA with the within-subject factor ‘format’ and the dependent variable ‘mean fixation duration’ or ‘mean duration of silent pauses’ was calculated to test hypothesis three. In addition we calculated correlation analyses between item difficulty and mean fixation duration and item difficulty and silent pauses.

Results

Table 1 shows the means and standard deviations for all outcome measures under investigation in this study. The results will be presented per hypothesis below.

Performance

The overall item difficulty differed between $P_{\min}=0.45$ and $P_{\max}=0.82$ with a mean item difficulty of $M=0.66$ ($SD=0.28$). Comparing items in the original and adapted format, we found a significant difference in item difficulty ($F(1,28)=10.48$, $p=0.03$, $\eta^2=.27$). Items in the original format were significantly more difficult indicated by a lower average probability score ($M=0.58$; $SD=0.32$) as compared to items in the adapted format ($M=0.74$; $SD=0.24$).

Visual search

Fixation duration

Before we tested our hypotheses, we first checked for differences in total time spent on the adapted and the original items. A one-way repeated-measures ANOVA showed a significant difference between the time spent on the two item formats ($F(1, 28)=5.40$, $p=0.03$, $\eta^2=0.16$).

Then we looked if the relative time spent on the areas of interest differed between the adapted and original items. A repeated measures MANOVA with the within-subject factor ‘format’ and the dependent variable ‘relative fixation duration’ was calculated. Results show a main effect of format on fixation duration ($F=(2, 27)=15.18$, $p<0.001$, $\eta^2=0.53$). Univariate tests show that format did lead to different viewing times on both the ‘items stimulus’ and ‘item stem’. Adapted formatting led to longer viewing times for

Table 1 Means and standard variations of all dependent variables

	<i>N</i>	Original items <i>M (SD)</i>	Adapted items <i>M (SD)</i>	<i>p</i> -value
Performance				
Item difficulty	33	.58 (.32)	.74 (.24)	.03*
Visual search				
Relative fixation duration item stem	29	13% (.05)	19.81% (.08)	<.001*
Relative fixation duration item stimulus	29	53.25% (.11)	47.24% (.12)	<.001*
Revisits item	29	10.87 (4.70)	17.34 (7.30)	<.001*
Revisits context	29	28.07 (11.99)	22.21 (9.31)	<.001*
Cognitive load				
Mean fixation duration (ms)	29	353.62 (79.18)	358.26 (91.51)	n.s.
Mean duration of silent pauses (sec)	29	33.23 (15.77)	25.99 (10.60)	<.001*

the AOI 'item stem' ($F=(1, 28)=28.99, p<0.001, \eta^2=0.51$ and shorter viewing times for the AOI 'item stimulus' ($F=(1, 28)=14.73, p<0.001, \eta^2=0.35$). The hypothesis that the adapted items would lead to longer fixation durations on the item stem and shorter fixation durations on the item stimulus than original items is hereby confirmed.

In post-hoc analyses we also analyzed fixation times for the items that were adapted in cluster one (4 items), cluster two (3 items) and cluster three (3 items). These analyses showed significant results for items in cluster one ($F=(2, 27)=10.77, p<0.001, \eta^2=0.44$) and two ($F=(2, 27)=7.79, p=0.002, \eta^2=0.37$) but not for items in cluster three ($F=(2, 27)=1.27, p=0.30, \eta^2=0.09$). Univariate tests showed significant higher fixation times on the 'item stem' AOIs and lower fixation time on the 'item stimulus' AOIs for items in cluster one ($F=(1, 28)=22.15, p<0.001, \eta^2=0.44$ and $F=(1, 28)=5.31, p=0.03, \eta^2=0.16$). For items in cluster two the same pattern evolved ($F=(1, 28)=14.881, p=0.001, \eta^2=0.35$ and ($F=(1, 28)=9.92, p=0.004, \eta^2=0.26$).

Revisits

A repeated measures MANOVA with the within-subject factors 'format' and the dependent variable 'number of revisits per AOI' was calculated. Results show a significant effect of format ($F(2, 27)=31.26, p<0.001, \eta^2=0.70$). Univariate tests show that the number of revisits to the 'item stem' AOIs of the adapted items is significantly greater than the number of revisits to the item stem AOIs of the originally formatted items, $F(1, 28)=38.18, p<0.001, \eta^2=0.58$. Moreover, in the adapted format there are significantly fewer revisits to the 'item stimulus' as compared to the original formatted items, $F(1, 28)=9.69, p<0.001, \eta^2=0.26$. It thus seems that adapting the items helped students in their visual search in the sense that it helped them to focus more on the 'item stem' (question and answer) and be less distracted by the 'item stimulus' (picture and explanatory information). The hypothesis that the adapted format leads to fewer

revisits to the explanatory and picture but more revisits to the question and answer was confirmed.

In post-hoc analyses we also analyzed revisits for the items in cluster one (4), cluster two (3 items) and cluster three (3 items). These analyses showed significant results for items in cluster one ($F=(2, 27)=19.08, p<0.001, \eta^2=0.59$) and two ($F=(2, 27)=16.73, p<0.001, \eta^2=0.57$) but not for items in cluster three ($F=(2, 27)=1.74, p=0.20, \eta^2=0.12$). Univariate tests showed for cluster one significant more revisits on the ‘item stem’ AOIs but not on the ‘item stimulus’ AOIs ($F=(1, 28)=21.20, p<0.001, \eta^2=0.43$ and ($F=(1, 28)=0.39, p=0.84, \eta^2=0.001$). For items in cluster two there were significantly more revisits to the ‘item stem’ AOIs and less revisits to the ‘item stimulus’ AOIs ($F=(1, 28)=12.28, p=0.002, \eta^2=0.32$ and ($F=(1, 28)=11.04, p=0.003, \eta^2=0.30$).

Cognitive load

Pearson correlation analysis showed only significant correlations between duration of silent pauses and item difficulty for original items ($p<0.002$). There were no significant correlations between duration of silent pauses and item difficulty in the adapted items nor between mean fixation duration and item difficulty for original nor adapted items.

Mean fixation duration

A one-way repeated measures ANOVA with the within-subject factor ‘format’ and the dependent variable ‘mean fixation duration on the item’ was calculated. Results showed no significant effect of format on mean fixation duration ($F=(1, 28)=0.27, p=0.61, \eta^2=0.01$). The hypothesis that the adapted items would lead to lower cognitive load was not confirmed.

In post-hoc analyses we also analyzed fixation durations for the items in cluster one (4), two (3 items) and three (3 items). These analyses showed no significant results for items in cluster one ($F=(1, 28)=6.72, p=0.021, \eta^2=0.19$) and two ($F=(1, 28)=5.99, p=0.02, \eta^2=0.18$) but only for items in cluster three ($F=(1, 28)=6.72, p=0.02, \eta^2=0.19$).

Silent pauses

A one-way repeated measures ANOVA with ‘format’ as within-subject variable and ‘duration of silent pauses’ as dependent variable shows a significant difference in average silent pauses between the two different formats with a large effect size ($F(1, 28)=15.92, p<0.001, \eta^2=0.36$). The mean duration of the silent pauses (in seconds) is larger for the originally formatted items ($M=33.29; SD=15.78$) as compared to the adapted items ($M=25.99; SD=10.60$). The hypotheses that the adapted items lead to lower cognitive load, as indicated by shorter silent pauses, as compared to original items is hereby confirmed.

In post-hoc analyses we also analyzed silent pauses for the items in cluster one (4), two (3 items) and three (3 items). These analyses showed significant results for items in cluster one ($F=(1, 28)=4.87, p=0.04, \eta^2=0.15$) and two ($F=(1, 28)=14.17, p=0.001, \eta^2=0.34$) but not for items in cluster three ($F=(1, 28)=1.68, p=0.21, \eta^2=0.06$).

Discussion

Computer-based testing provides the opportunity to include multimedia material (i.e., pictures, audio, etc.). However, the way such items are designed very much affects how students perform on them (e.g., Ögren et al., 2016). In this study, we investigated the effects of adapting computer-based test items according to the principles of multimedia learning on performance (i.e., operationalized as mean item difficulty) and used process measures to be able to better understand how the design affects performance. For that purpose, items from a standardized computer-based test were converted. Based on prior research, we expected that adapted items were less difficult as compared to the original items (cf. Lindner et al., 2017a; *performance hypothesis*). From a process perspective, we furthermore expected that adapting items would lead to more beneficial visual processing (cf. Jarodzka et al., 2015; Ögren et al., 2016; *visual search hypothesis*) and lower cognitive load (i.e., shorter fixation durations and less silent pauses) (cf. CTML; *cognitive load hypothesis*).

The results confirmed our *performance hypothesis*. Adapting items according to the CTML and CLT lowers item difficulty and thus items in the adapted format are easier for students. This result is in line with earlier studies (Lindner et al., 2016; Lindner, 2017).

Second, in accordance with the *visual search hypothesis*, the results showed that adapting items according to the principles of CTML led to longer viewing times on the question and answer but shorter viewing times on the pictures and explanatory information. These results suggest that applying multimedia principles to test items, leads to a more balanced viewing behavior in which students pay more attention to the question and answer sections and are less distracted by visual elements. In addition, in the adapted items there were fewer revisits to pictures and explanatory information and more revisits to question and answer parts of the items indicating more integration between the different elements.

Finally, the results partly confirmed our *cognitive load hypothesis*. Adapting items according to the principles of CTML and CLT did not result in lower overall cognitive load as indicated by shorter mean fixation durations but it did result in shorter silent pauses. Prior research has shown that more objective measures such as fixation durations and dual-task performance are not always useful for measuring extraneous load (van Gog et al., 2009). The here used stimuli contained a substantial amount of text, which might have affected results (Rayner, 2009). In hindsight, fixation duration might not have been the optimally sensitive measure for such text-rich stimuli. For future research, different measures such as the mental effort rating scale (Paas, 1992) might be more useful.

Taking these results all together, adapting multimedia computer-based items to conform to the multimedia principles enables students to increase their success in processing and comprehending the test questions and in turn, increasing success in answering the questions. Processing data indicates that this is probably because there is more attention to the question and answer and more integration of the different elements (i.e., question, explanatory text, answer, picture). Previous research has shown that students often do not read the question text carefully, which probably results in them making mistakes when solving the test item. (e.g., Bully & Valencia, 2002). Our findings are in line with such previous findings. Eliminating decorative pictures, redundant information, and integrating text-picture material seems to draw students' attention towards the question and answer (see also Jarodzka et al., 2015) and makes it easier to process the contextual information resulting in lower item difficulty (i.e., better performance). One possible explanation for that might be that students understand the questions better because they read it more carefully and

integrate the information better but more research is needed to be able to explain effects properly.

Post-hoc analyses, however, also reveal that items for which all visual elements were removed, did not benefit from the application of the multimedia principles. Students thus seem to benefit when (relevant) visual elements are used in test-items (i.e., multimedia-effect) in the right way. Maybe because it increases comprehension by compensating for low reading ability (Wallen et al., 2005) and/or it enhances student interest and attention for the item (Wang & Adesope, 2014). Future research should more closely look into the effects of visualizations on item validity, difficulty, and processing.

Limitations

In the present study, we used a standardized test and adapted its items in such a way that they were in line with the principles in CTML and CLT. Our experimental approach of using authentic tests has the strong advantage that our findings are ecologically valid and thus are highly for educational practice. On the downside, this study did not examine how each of the single multimedia guidelines affect the performance and processing of computer-based test items when used in the absence of other multimedia guidelines, nor did this study examine and compare the relative impact of each guideline when used simultaneously and in conjunction with other guidelines. More research is thus needed to investigate under which conditions applying multimedia guidelines benefit CBT and under which conditions they hamper CBT. Related to this issue is the question how the multimedia design affects test performance. It can first affect students' understanding of the test questions, or it can provide better retrieval cues in for example the answer options (Kirschner et al., 2016). These are two different research lines which might be combined in future research.

A second limitation of our study is that we focused on one domain only, namely mathematics and the sample has generally low verbal abilities. If we used another domain such as language learning and/or conducted our studies among a different sample, our results might have been different. Especially since in some domains, it is rather important that students can select the right/useful information from a wider set of (irrelevant) information causing a shift in the purpose of the assessment. For future research, it is thus important to investigate the design principles for testing in different domains, among different populations, and for different test goals.

Third, we used eye-tracking and silent pauses to unravel cognitive processes. Both methods have their drawbacks. For instance, dual-task performance may slow down primary task performance, put additional load on participants, or be incomplete (in particular for highly visual processes). Still, dual-task performance is a well-established method to study cognitive processes ever since the widely-cited work of Ericsson and Simon (1993). Likewise, eye-tracking has its own disadvantages: it can only tell us, where someone looked at but not *why*. Still, it allows us to see, *which* information enters the cognitive system, in which order, and which elements are attended to for how long. In that way, both methods very well complement each other to come to a more complete picture of the cognitive processes at hand (e.g., Helle, 2017).

Conclusion and educational relevance

This study is one of few studies investigating the effect of applying several multimedia principles to testing. Despite the limitations, the current study shows promising results of applying multimedia principles of learning to testing. In general, the results show that the central idea underlying CTML and CLT is also applicable to CBT (Kirschner et al., 2016). More specifically, the results seem to indicate that when designing computer-based tests, it is important to reduce extraneous load by integrating text and pictures and delete seductive elements as much as possible. However, post-hoc analyses also showed that items from which relevant visualizations were removed, did not benefit from the redesign.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval It is also approved by our own universities ethical committee (reference number: U2017/03387/FRO).

Informed consent Informed consent was obtained from all individual participants included in the study.

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

- Abercrombie, S. (2013). Transfer effects of adding seductive details to case-based instruction. *Contemporary Educational Psychology*, 38, 149–157. <https://doi.org/10.1016/j.cedpsych.2013.01.002>
- Alemdag, E., & Cagiltay, K. (2018). A systematic review of eye-tracking research on multimedia learning. *Computer and Educations*, 125, 413–428. <https://doi.org/10.1016/j.compedu.2018.06.023>
- Anglin, G. J., Vaez, H., & Cunningham, K. L. (2004). Visual Representations and Learning: The Role of Static and Animated Graphics. In D. H. Jonassen (Ed.), *Handbook of research on educational communications and technology* (pp. 865–916). Lawrence Erlbaum Associates Publishers.
- Azabdaftari, B., & Mozaheb, M. A. (2012). Comparing vocabulary learning of EFL learners by using two different strategies: Mobile learning vs. flashcards. *The Eurocall Review*, 20(2), 47–59.
- Baddeley, A. (1992). Working memory. *Science*, 255(5044), 556–559. <https://doi.org/10.1126/science.1736359>
- Basu, A., Cheng, I., Prasad, M., & Rao, G. (2007). *Multimedia adaptive computer based testing: An overview*. Multimedia and Expo, 2007 IEEE International Conference on. <https://doi.org/10.1109/ICME.2007.4285034>.
- Başoğlu, E. B., & Akdemir, Ö. (2010). A comparison of undergraduate students' English vocabulary learning: Using mobile phones and flashcards. *TOJET: the Turkish Online Journal of Educational Technology*, 9(3), 1–7.
- Beddow, P. A. (2018). Cognitive load theory for test design. In S. N. Elliott, R. J. Kettler, P. A. Beddow, & A. Kurz (Eds.), *Handbook of accessible instruction and testing practices. Issues, innovations and applications* (2nd ed., pp. 199–212). Springer International Publishing.

- Beddow, P. A., Elliott, S. N., & Kettler, R. J. (2009). *Accessibility rating matrix*. Retrieved on July 2018 from https://peabody.vanderbilt.edu/docs/pdf/PRO/TAMI_Accessibility_Rating_Matrix.pdf
- Bully, M. R., & Valencia, S. W. (2002). Below the bar: Profiles of students who fail state reading assessments. *Educational Evaluation and Policy Analysis, 24*(3), 219–239. <https://doi.org/10.3102/01623737024003219>
- Butcher, K. R. (2014). The multimedia principle. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (2nd ed., pp. 174–205). Cambridge University Press.
- Brünken, R., Plass, J. L., & Leutner, D. (2003). Assessment of cognitive load in multimedia learning with dual-task methodology: Auditory load and modality effects. *Instructional Science, 31*(2), 103–120. <https://doi.org/10.1023/B:TRUC.0000021812.96911.c5>
- Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction, 8*(4), 293–332. https://doi.org/10.1207/s1532690xci0804_2
- Choi, H.-H., van Merriënboer, J. J. G., & Paas, F. (2014). Effects of the physical environment on cognitive load and learning: Towards a new model of cognitive load. *Educational Psychology Review, 26*(2), 225–244. <https://doi.org/10.1007/s10648-014-9262-6>
- Chua, Y. P., & Don, Z. M. (2013). Effects of computer-based educational achievement test on test performance and test takers' motivation. *Computers in Human Behavior, 29*(5), 1889–1895. <https://doi.org/10.1016/j.chb.2013.03.008>
- Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. MIT Press.
- Helle, L. (2017). Prospects and pitfalls in combining eye-tracking data and verbal reports. *Frontline Learning Research, 5*(3), 81–93.
- Hyövä, J. (2010). The use of eye movements in the study of multimedia learning. *Learning and Instruction, 20*, 172–176. <https://doi.org/10.1016/j.learninstruc.2009.02.013>
- Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., van de Jarodzka, H., & Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. Oxford: Oxford University Press.
- Jarodzka, H., Janssen, N., Kirschner, P. A., & Erkens, G. (2015). Avoiding split attention in computer-based testing: Is neglecting additional information facilitative? *British Journal of Educational Technology, 46*(4), 803–817. <https://doi.org/10.1111/bjjet.1217433>
- 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>
- Kaplan-Rakowski, R., & Loranc-Paszyk, B. (2017). Students' views on the helpfulness of multimedia components of digital flashcards in mobile-assisted vocabulary learning. In K. Borthwick, L. Bradley, & S. Thouésny (Eds.), *CALL in a climate of change: Adapting to turbulent global conditions - short papers from EUROCALL 2017* (pp. 170–176). Research-publishing.net.
- Kirschner, P. A., Park, B., Malone, S., & Jarodzka, H. (2016). Toward a cognitive theory of multimedia assessment (CTMMA). In M. J. Spector, B. B. Lockee, & M. D. Childress (Eds.), *Learning, design, and technology: An international compendium of theory, research, practice, and policy* (pp. 1–23). Springer International Publishing.
- Korbach, A., Brünken, R., & Park, B. (2017). 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>
- Korbach, A., & Brünken, R. (2017). Measurement of cognitive load in multimedia learning: A comparison of different objective measures. *Instructional Science, 45*(4), 515–536. <https://doi.org/10.1007/s11251-017-9413-5>
- Lin, Y. Y., Holmqvist, K., Miyoshi, K., & Ashida, H. (2017). Effects of detailed illustrations on science learning: An eye-tracking study. *Instructional Science, 45*, 557–581. <https://doi.org/10.1007/s11251-017-9417-1>
- Lin, C. C., & Yu, Y. C. (2017). Effects of presentation modes on mobile-assisted vocabulary learning and cognitive load. *Interactive Learning Environments, 25*(4), 528–542. <https://doi.org/10.1080/10494820.2016.1155160>
- Lindner, M. A., Ihme, J. M., Saß, S., & Köller, O. (2016). How representational pictures enhance students' performance and test-taking pleasure in low-stakes assessment. *European Journal of Psychological Assessment, 31*(1), 1–10. <https://doi.org/10.1027/1015-5759/a000351>
- Lindner, M. A., Eitel, A., Strobel, B., & Köller, O. (2017). Identifying processes underlying the multimedia effect in testing: An eye-movement analysis. *Learning and Instruction, 47*, 91–102. <https://doi.org/10.1016/j.learninstruc.2016.10.007>
- Mayer, R. E. (2009). *Multimedia learning* (2nd ed.). Cambridge University Press.
- Mayer, R. E. (2014a). Cognitive theory of multimedia learning. In R. E. Mayer (Ed.), *Cambridge handbook of multimedia learning (Cambridge Handbooks in Psychology)* (pp. 43–71). Cambridge University Press.

- Mayer, R. E. (2014b). Principles for reducing extraneous processing in multimedia learning: Coherence, signaling, spatial contiguity, and temporal contiguity principles. *Cambridge handbook of multimedia learning (Cambridge Handbooks in Psychology)* (pp. 279–315). Cambridge University Press.
- Mayer, R. E. (2019). Taking a new look at seductive details. *Applied Cognitive Psychology*, 33(1), 139–141. <https://doi.org/10.1002/acp.3503>
- Ögren, M., Nyström, M., & Jarodzka, H. (2016). There's more to the multimedia effect than meets the eye: Is seeing pictures believing? *Instructional Science*, 44(5), 1–25. <https://doi.org/10.1007/s11251-016-9397-6>
- Paivio, A. (1986). *Mental representations*. New York: Oxford University Press.
- Paas, F. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive load approach. *Journal of Educational Psychology*, 84, 429–434.
- Paas, F., & Sweller, J. (2014). Implications of cognitive load theory for multimedia learning. *The Cambridge Handbook of Multimedia Learning*. <https://doi.org/10.1017/cbo9781139547369.004>
- Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3), 372–422. <https://doi.org/10.1037//0033-2909.124.3.372>
- Rayner, K. (2009). The 35th Sir Frederick Bartlett Lecture: Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology*, 62(8), 1457–1506. <https://doi.org/10.1080/17470210902816461>
- Rijksoverheid (2018). Referentieniveaus taal en rekenen [website]. Retrieved on 6th Juli 2018 from <https://www.rijksoverheid.nl/onderwerpen/taal-en-rekenen/referentiekader-taal-en-rekenen>
- Sweller, J. (2011). Cognitive load theory. In J. P. Mestre & B. H. Ross (Eds.), *The psychology of learning and motivation: Vol. 55. The psychology of learning and motivation: Cognition in education* (pp. 37–76). Elsevier Academic Press.
- Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive load theory* (Vol. 1). Springer.
- Sweller, J., Van Merriënboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, 10(3), 251–296. <https://doi.org/10.1023/A:102219372820535>
- Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review*, 31(2), 261–292. <https://doi.org/10.1007/s10648-019-09465-5>
- Underwood, G., Jebbett, L., & Roberts, K. (2004). Inspecting pictures for information to verify a sentence: Eye movements in general encoding and in focused search. *The Quarterly Journal of Experimental Psychology Section A*, 57, 165–182. <https://doi.org/10.1080/02724980343000189>
- van Gog, T., Kester, L., Nievelstein, F., Giesbers, B., & Paas, F. (2009). Uncovering cognitive processes: Different techniques that can contribute to cognitive load research and instruction. *Computers and Human Behavior*, 25, 325–331. <https://doi.org/10.1016/j.chb.2008.12.021>
- van Gog, T., Paas, F., van Merriënboer, J., & Witte, P. (2005). Uncovering the problem-solving process: Cued retrospective reporting versus concurrent and retrospective reporting. *Journal of Experimental Psychology: Applied*, 11, 237–244. <https://doi.org/10.1037/1076-898X.11.4.237>
- van Gog, T., & Jarodzka, H. (2013). Eye tracking as a tool to study and enhance cognitive and metacognitive processes in computer-based learning environments. In R. Azevedo & V. Aleven (Eds.), *International handbook of metacognition and learning technologies. Springer international handbooks of education* (pp. 143–156). Springer.
- Wallen, E., Plass, J. L., & Brünken, R. (2005). The function of annotations in the comprehension of scientific texts: Cognitive load effects and the impact of verbal ability. *Educational Technology Research and Development*, 53(3), 59–71.
- Wang, Z., & Adesope, O. O. (2014). Effects of seductive details on multimedia learning. *Journal of Studies in Education*, 4, 32–44. <https://doi.org/10.5296/jse.v4i3.6024>
- Yin, B., & Chen, F. (2007). Towards automatic cognitive load measurement from speech analysis. In J. A. Jacko (Ed.), *Human-computer interaction. Interaction design and usability. HCI 2007. Lecture notes in computer science*, 4550. Springer. https://doi.org/10.1007/978-3-540-73105-4_111