British Journal of Educational Technology *Vol 46 No 4 2015* 803–817
doi:10.1111/bjet.12174

Avoiding split attention in computer-based testing: Is neglecting additional information facilitative?

Halszka Jarodzka, Noortje Janssen, Paul A. Kirschner and Gijsbert Erkens

Halszka Jarodzka is an assistant professor at the Center for Learning Sciences and Technologies at the Open University of the Netherlands. Her research interests lie in the use of eye tracking to investigate processes relevant for educational psychology. Her main focus is in characteristics of visual expertise and its training by eye movement modeling examples. Noortje Janssen is a PhD student at the Department of Instructional Technology of the University of Twente. Her research interests lie in supporting teachers in implementing information and communication technology in their classroom practices. Paul A. Kirschner is full professor of Educational Psychology and head of research at the Center for Learning Sciences and Technologies at the Open University of the Netherlands. He is acknowledged expert in the field of computer-supported collaborative learning, instructional design, and the use information and communication technologies for educational and instructional purposes. Gijsbert Erkens is researcher in educational psychology and computer-supported collaborative learning at the Department of Education, Faculty of the Social and Behavioral Sciences, Utrecht University. Address for correspondence: Dr Halszka Jarodzka, Welten Institute Research Centre for Learning, Teaching and Technology, Open University of the Netherlands, P.O. Box 2960, 6401 DL Heerlen, The Netherlands. Email: Halszka.Jarodzka@OU.nl

Abstract

This study investigated whether design guidelines for computer-based learning can be applied to computer-based testing (CBT). Twenty-two students completed a CBT exam with half of the questions presented in a split-screen format that was analogous to the original paper-and-pencil version and half in an integrated format. Results show that students attended to all information in the integrated format while ignoring information in the split format. Interestingly, and contrary to expectations, they worked more efficiently in the split format. A content analysis of the ignored information revealed that it was mostly not relevant to answering the questions, unnecessarily taxed students' cognitive capacity and inefficiently increased the mental effort they expended. Further comparisons of different mental effort measures indicate that mental effort had an explicit (ie, self-reports, explicit utterances) and an implicit component (ie, silent pauses in thinking-aloud, eye tracking parameters). Consequently, when designing CBT environments, not only the design of the tasks but also the content of the given information and their effect on the different aspects of mental effort must be considered.

Introduction

As technology use in learning increases, the use of technology for assessment is also becoming more common. In the Netherlands, for example, computer-based testing (CBT) is making its way into the national exams used in secondary education (De Boer, 2009) by the Dutch Institute of Educational Measurement (CITO; www.cito.nl) that develops standardized tests for schools on behalf of the government. CBT provides the possibility to include a mix of different presentation formats, such as videos, text, pictures, etc. (ie, multimedia: Mayer, 2005a), which is increasingly being implemented (cf. Parshall, Harmes, Davey & Pashley, 2010). Researchers, however, have emphasized that it is important to appropriately *design* CBT environments so that students can focus on test content and are not impeded by difficulties relating to its design. For instance, Parshall, Spray, Kalohn and Davey (2002, p 5) stated that ". . . examinees need to know clearly to what part of the screen they must attend to, how to navigate, and how to indicate a response.

Practitioner Notes

What is already known about this topic

- Learning material presented in a split format (ie, related information is either spatially or temporally separated) causes unnecessary visual search for the learner.
- Unnecessary visual search for information requires excessive expenditure of cognitive capacity (ie, mental effort).
- These capacities are no longer available for learning, hence, splitting the learners' attention hampers learning.

What this paper adds

- Split-attention effect in testing.
- Directly investigating the amount of visual search in split versus integrated formats with eye tracking.
- Triangulation of different measures of mental effort reveals two aspects of cognitive load: implicit and explicit.

Implications for practice and/or policy

- While the content of testing material is central, it is important to design its layout based on cognitive theories.
- Instructional design guidelines for learning cannot be directly translated to testing, because these differ too much from each other.
- In testing, the relevancy of given information needed for answering the test questions should be considered explicitly in relation to the skills that are meant to be measured (ie, the aim of testing).
- Multimedia can be a valuable addition to testing environments if the test is meant to capture the students' ability to filter relevant out of irrelevant information.

[. . .] The more 'intuitive' the computer test software is, the less the examinee needs to attend to it, rather than to the test questions . . ." (p 5).

Thus, though the technical possibilities to implement multimedia in testing exist, the question of how such multimedia CBT should look from a pedagogical perspective is open to discussion. As there are yet neither theories nor empirical guidelines on this specific topic, we tend to make use of what we do have, namely models and design guidelines for *learning* with multimedia. Though there are important differences between learning and testing with multimedia, there are also commonalities, not the least of which is that both aim first at *understanding* the multimedia materials. According to well established multimedia learning theories (ie, Cognitive Theory of Multimedia Learning: Mayer, 2001; Cognitive Load Theory: Sweller, Van Merriënboer & Paas, 1998), the prerequisites for this understanding build upon general assumptions of human cognitive architecture. The assumption is that the human cognitive system is limited in capacity with respect to how much new information it can process at any one time (not in long-term memory storage, though). This dates back to early research on the structure and functioning of human working memory (eg, Baddeley, 1992; Miller, 1956). Hence, this assumption holds not only for learning, but also for task performance, such as in testing. The second assumption is that information should be actively processed to be understood. This is based on Atkinson and Shiffrin's (1968) information processing model, which again is not a specific learning model, but instead describes general information processing. Therefore, we argue that the theories on learning with

multimedia *could* also be applied to information processing in testing, but of course should be tested (as we do here).

These theories on multimedia learning aim at reducing unnecessary mental effort caused by poor instructional design which hampers learning (Cognitive Theory of Multimedia Learning: Mayer, 2005b; Cognitive Load Theory: Sweller *et al*, 1998). One of the guidelines to facilitate learning from multimedia is to avoid the *split attention effect* (Chandler & Sweller, 1991; also called spatial contiguity principle; Moreno & Mayer, 1999).

Split attention in a computer-based environment occurs when information that needs to be integrated is divided either temporally between several successive screens or spatially across one screen. Such distributed presentations require learners' to *visually search* for related information that needs to be integrated and constantly shift attention from one information element to another. The difficulty lies in keeping the first information element active in working memory, while the other is looked for and attended to so that both can be integrated. Hence, a 'split format' requires unnecessary search processes that consume cognitive capacity and cause mental effort. This capacity is no longer available for learning, which in turn results in lower performance. An integrated format has been shown to be more efficient for learning compared to a split format (Ward & Sweller, 1990). It is important to note, though, that 'split' and 'integrated' formats are not two distinct poles, but rather a continuum; hence, information can be presented more or less split. For instance, information can be split across several pages causing not only spatial, but also temporal discontinuity (cf. Mayer, 2001). For the purpose of the current study we refer to the definition of Sweller *et al* (1998) of a split (all information is presented on two parts of the same page) and an integrated format (supplementary information is presented within the text, right when it is needed).

This study investigated whether the split-attention effect also holds true for *testing*. When doing this, we further need to examine the two main assumptions underlying this effect: unnecessary visual search and increased mental effort and discuss how both can be measured.

Unnecessary visual search in a split format investigated by means of eye tracking

To investigate unnecessary visual search for related information, it is important to actually measure these processes, for example through eye tracking. Eye tracking reveals what a person looks at, for how long, and in which order (Holmqvist *et al*, 2011). As looking at certain elements is closely related to cognitively processing these elements, eye tracking captures visual and cognitive aspects of attention (Just & Carpenter, 1976).

With this method, researchers found that, when learning from text and pictures, learning is heavily driven by the text while pictures are only minimally inspected (Hannus & Hyönä, 1999). Furthermore, pictures only improved learning of those text passages that were also illustrated in the picture and not the remaining text passages. While both high- and low-ability students often switch their visual focus between both representations (so-called *transitions*), indicating that when learning from text and pictures both sources need to be integrated simultaneously (Hegarty & Just, 1993), low-ability students made many transitions indicating that they had difficulties integrating information from the two sources. These transitions between the text and the picture are indicators for a large amount of visual search. Holsanova, Holmberg and Holmqvist (2008) investigated the effect of a split versus an integrated format in a naturalistic newspaper reading study. They found more transitions between semantically related text and picture parts in an integrated format compared to a split format indicating that visual search and integration might not even take place in a split format.

Thus, eye tracking allows studying visual search underlying a split or an integrated format and shows that integration of related information is helpful, a split format, however, might prevent this.

Increased mental effort in a split format

A further assumption of the split attention effect is the increase of mental effort. Cerpa, Chandler and Sweller (1996) showed that in learning to use software an integrated format lead to better learning results than a split format, because it required less perceived mental effort as self-reported by the participants. Similar relations have been shown in the domain of chemistry (Kalyuga, Chandler & Sweller, 1999). These studies all used a 1-item self-report scale (Paas, 1992) to determine mental effort expended.

As this approach has been criticized for various reasons (eg, De Jong, 2010) more objective means of measuring mental effort have been developed (for an overview on the different measures see Van Mierlo, Jarodzka, Kirschner & Kirschner, 2012). In the human-computer-interaction community, indicators for cognitive load in *thinking-aloud* have become a topic of interest (eg, Ericsson & Simon, 1993; Müller, Grossmann-Hutter, Jameson, Rummer & Wittig, 2001). Yin and Chen (2007), for instance, showed that pauses in thinking-aloud are a strong indicator for cognitive overload. Furthermore, think-aloud protocols could be coded in terms of whether they are explicitly indicative of extraneous load (eg, statements on the high difficulty of the task or the difficulty of finding the necessary information). In that way, two aspects of cognitive load could be captured, namely the overall mental effort evidenced by pauses in thinking-aloud (ie, silence) and the amount of experienced harmful (ie, extraneous) mental effort by such utterances. To our knowledge, this methodology has not yet been used to investigate mental effort underlying the split-attention effect.

Another option are physiological measures of cognitive load based on the assumption that there is a relationship between changes in mental effort and changes in physiological states. One non-intrusive example of physiological measures is *eye tracking* (Holmqvist *et al*, 2011). Higher mental effort has been shown to be related to increased pupil dilation (eg, Klingner, Tversky & Hanrahan, 2011) and decreased fixation durations (eg, Van Orden, Limbert, Makeig & Jung, 2001). A study by Underwood, Jebbett and Roberts (2004) can be related to the split-attention effect. The authors showed participants a sentence either below a photograph (ie, integrated format) or on a separate page after a photograph disappeared (ie, temporally split format).They found that fixation durations were shorter when the text was presented with the picture than when it was presented afterwards. Though not specifically aimed at studying split-attention, the data points in the direction that a temporally split format increased fixation durations.

The question that arises here is not only whether there is a observable eye-tracked split-attention effect, but also whether the 'rules' for learning materials derived from cognitive load theory (Sweller *et al*, 1998) and the cognitive theory of multimedia learning (Mayer, 2005b) can be applied to CBT or whether a *cognitive theory of multimedia testing* is necessary.

Present study: hypotheses

Based on the request of the Dutch Institute of Educational Measurement (CITO) to improve and evaluate their current design of the testing environment, the current study applies the aforementioned information integration principle (ie, avoiding split attention) to *testing*. Comparable to learning, we may assume that *cognitive test load* can be induced by the complexity of the testing task itself and/or inadequate design of the testing environment. Presenting all information required to solve the test task in an integrated manner should minimize the need for the testee to utilize search processes for the corresponding information. Thus, an integrated format should lead to fewer transitions between the different information sources (Hypothesis 1).

To gain a deeper insight into the visual processes, researchers often compare the time spent on the pictorial and the textual information. However, as the findings are inconsistent and are very likely to depend on the amount of textual information (cf. Hannus & Hyönä, 1999; Underwood *et al*,

2004) an open research question was formulated on whether textual or pictorial information is inspected longer (Research Question 1).

Less visual search should allow as much cognitive capacity as possible to solve the testing task (Hypothesis 2). To validly capture the effects of format design on experienced test load different techniques will be triangulated. The first measure is a *subjective rating* of perceived mental effort. The second is *thinking-aloud*, where mental overload has been shown to be reflected in silent pauses. Also, the verbal data coming from the thinking-aloud protocols is coded in terms of whether hints for hindered mental effort were uttered. The third measure of mental effort is operationalized through the capture of *eye tracking* parameters, such as pupil dilations and fixation durations. As an additional research question, this research investigates how these different mental effort measures relate to each other (Research Question 2).

Thus, presenting a testing task in an integrated format should result in more reliable testing outcomes (ie, those with the necessary knowledge and skills will also be able to successfully solve the testing task) while investing less mental effort in less time than in a split format. Hence, the final hypothesis is that an integrated testing format is more efficient than a split testing format (Hypothesis 3).

In the current study all participants were thoroughly trained for this experience (for more details see below).

Method

Participants

Twenty-seven pre-university students were recruited for this study. The exam used in this study was actually designed to be completed by students at the end of their 5th year of senior general education. Because the experiment was carried out at the beginning of the year, students on senior general education had not yet reached the level of exam and thus, we chose to use pre-university students which is a higher level (both in quality and quantity of learning). As such, we could be sure that the level of the items was appropriate. Five students were excluded from the analysis because of inaccuracy in the eye tracking data, resulting in valid data from 22 participants (1 male, 21 females; $M = 16.36$ years, $SD = 0.49$). Participation was voluntarily and was rewarded with a €5.00 gift voucher after the experiment.

Material and apparatus

CBT testing environment

The testing material is an authentic standardized national *Art Appreciation* exam for the Dutch secondary education as it was used nation-wide in 2010. This exam was designed and standardized by a consortium of educational practitioners (ie, that national testing body of the Netherlands that provides all final exams to all students in the Netherlands) to be used at the end of senior general education (and not just by the researchers for the current study). At this stage, students are supposed to have acquired several domain-specific content skills (eg, being able to explain how science and arts influence each other), but they also have to be able "adequately handle source material when reflecting" (College voor Examens (CEVO, 2010). The official exam we used in this study is supposed to capture all these skills.

The exam was presented in the software program ExamenTester®, currently used for the national CBT in the Netherlands (De Boer, 2009). Both test performance data and subjective mental effort ratings (see below) were logged with this software. The test items were—by default—presented in a split-screen format (Figure 1). In this study, an additional integrated format of the items was designed (Figure 2). Each item was composed of three components. The first was an explanatory text providing background for the test item (eg, "Courbet was not only progressive in his way of working. His view about the content of the art of painting was also novel as you will read in the

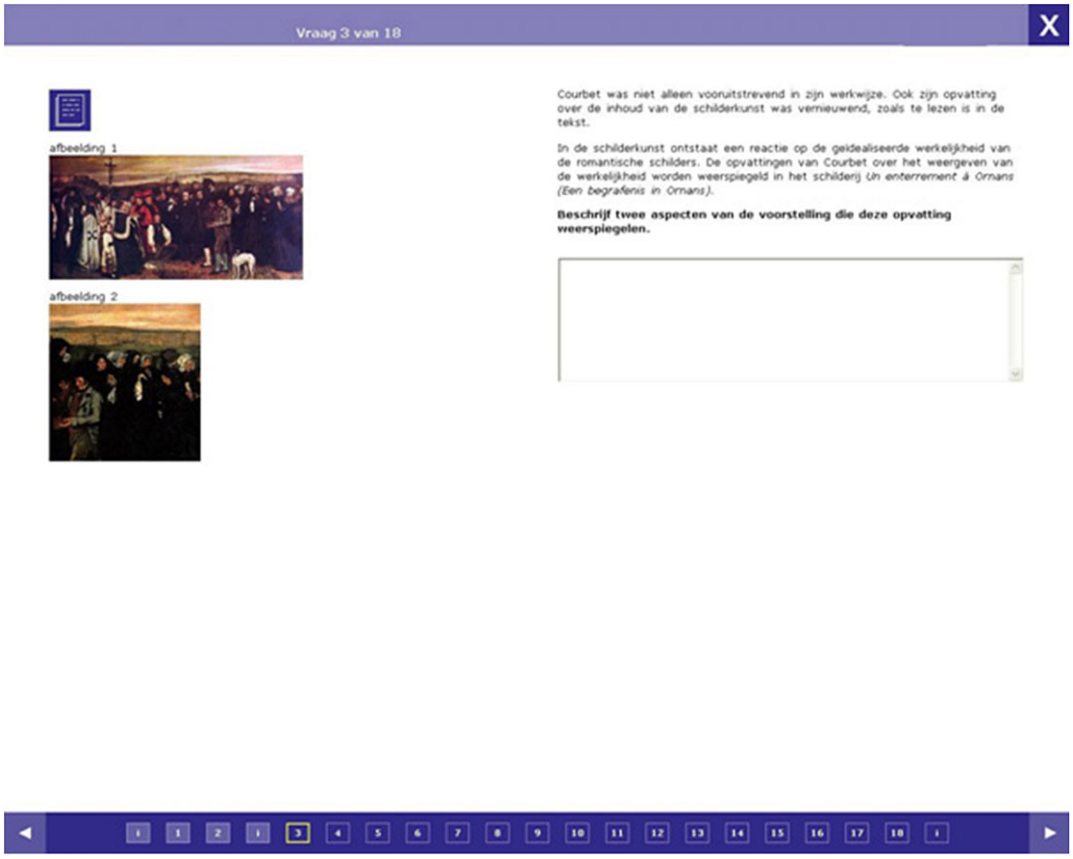

Figure 1: The original split format of the test questions used in this study. The task explanatory text is on the right-hand side, as well as the placeholder to fill in the answer. On the left-hand side the additional information sources can be found. Clicking on the text icon opens a pop-up window with additional textual information. Clicking on the small pictures opens a pop-up window with the same picture enlarged

text. In painting, a counter movement to the idealized reality of the romantic painters arose. Courbet's view on the reproduction of reality can be seen in the painting *Un enterrement à Ornans*."). The second component were clickable icons for additional information in pop-up windows (eg, text and painting in full and in detail view referred to in the explanatory text). These pop-ups inevitably introduced to some degree splitting attention for both formats. Of the eight items, six provided one additional textual information element and two provided two additional elements. Also, six out of eight items provided one additional pictorial information element (either a picture or a video), one provided two additional pictorial information elements, and one provided three. The final component was the test question presented in bold type together with a placeholder to type in an open format answer (eg, "Please describe two aspects of this painting that reflect this view").

Eye tracking equipment

Eye movements were recorded with a Tobii 1750 remote eye tracking system with temporal resolution of 50 Hz (Tobii, 2003, 2003), and analyzed with Tobii Studio 2.2.4 software (Tobii Studio, 2007, www.tobii.com).

Figure 2: The modified integrated format of the test questions used in this study. The task explanatory text is in the middle of the screen. The additional information sources are placed within the text, when they are referred to. The placeholder to fill in the answer is located at the bottom of the screen. Clicking on the text icon opens a pop-up window with additional textual information. Clicking on the small pictures opens a pop-up window with the same picture enlarged

Thinking aloud

Participants were trained and instructed in thinking aloud according to Ericsson and Simon (1993). They were asked to "verbalize everything that comes to mind, and to disregard the experimenter's presence in doing so, even if expletives were used (Van Gog, Paas, van Merriënboer & Witte, 2005). If they were silent for 5 seconds, they were reminded to keep thinking aloud (Van Gog *et al*, 2005). When a video with sound was running, participants were not reminded to think aloud, as they had to listen. The verbal data were recorded using a standard microphone attached to the stimulus PC.

Subjective mental effort rating

Participants were asked to rate their perceived mental effort after each test item ("How much effort did you invest to complete this task?", Paas, 1992) on a 9-point rating scale ranging from "very, very, very low effort" to "very, very, very high effort".

Procedure

The study was run in individual session of approximately 60 minutes. Participants were first trained in thinking aloud. Next, the eye tracking equipment was calibrated to the individual features of each participant based on a nine-point calibration protocol. Before the actual experiment started, participants completed a questionnaire to gather demographic data and then received a warm-up item for which no data were logged. Then, participants were instructed to

think aloud and complete eight test items as if it was a real exam. The original exam instructions, which did not specify whether to integrate information from different sources or whether to filter relevant from irrelevant information, were also used here. After each item, participants filled in the subjective mental effort rating. In the test, the split and integrated items were presented in an alternating order. Hence, each participant received four test items in a split format and four in an integrated format (ie, within-subjects design).

Data analysis

Scoring the test outcomes

The answers were scored by members of the Dutch Institute of Educational Measurement according to the same guidelines applied the year before when this test was used in Dutch schools. The values are presented in percentage correct.

Eye tracking analysis

The eye tracking recordings provided two types of dependent variables, namely visual search measures and mental effort measures. The data was filtered with the Tobii ClearView fixation filter, whereby a fixation definition of 30 pixels and 100 milliseconds was chosen (cf. Hegarty & Just, 1993; Loftus, 1981 for other materials including pictures). All analyses were performed with the Tobii Studio software version 2.2.4 (Tobii Studio, 2007).

To analyze *visual search*, all eye tracking parameters were assigned to certain on-screen elements (ie, areas of interest (AOIs). Three types of AOIs were defined, namely "explanatory text", "question & answer", and "additional information" (see Figure 3). The "additional information" AOI was further divided into textual information and pictorial information (ie, pictures or videos). As all additional information elements could be enlarged in a pop-up window, the individual recordings had to be divided into sections without pop-ups and in sections where pop-ups were enlarged. In each of these sections, the AOIs were defined accordingly. Two measures were derived from the AOI analysis. First, the total duration of participants' eye fixation per AOI was calculated to investigate which areas were most attended to in which presentation format. Second, the number of transitions between AOIs in terms of movements from one AOI to another AOI was calculated to investigate the number of comparisons between the different elements on the screen.

To capture *mental effort* experienced per test item, two eye tracking parameters were obtained: pupil dilation and fixation duration. For both parameters the mean per item and per participant was calculated.

Analysis of thinking-aloud protocols

Two different variables for mental effort were derived from the verbal data. First, silent pauses >2 seconds were counted per test item and coded as indicators of cognitive overload. Second, explicit utterances of elevated cognitive load during thinking-aloud (eg, "I wouldn't know", or "I think this is a difficult question") were counted.

Calculation of efficiency measures

Test outcomes were converted into standardized efficiency measure (Tuovinen & Paas, 2004). Therefore, the mean standardized mental effort score (z_{ME}) and the mean standardized time on task score (z_{tot}) were subtracted from the mean standardized test performance score (z_{perf}). The result was divided by the square root of 3[1]:

$$\frac{z_{perf} - z_{ME} - z_{tot}}{\sqrt{3}}$$

1. The reader is referred to Tuovinen and Paas (2004) for an explanation and motivation of this formula in determining the efficiency in *learning* conditions.

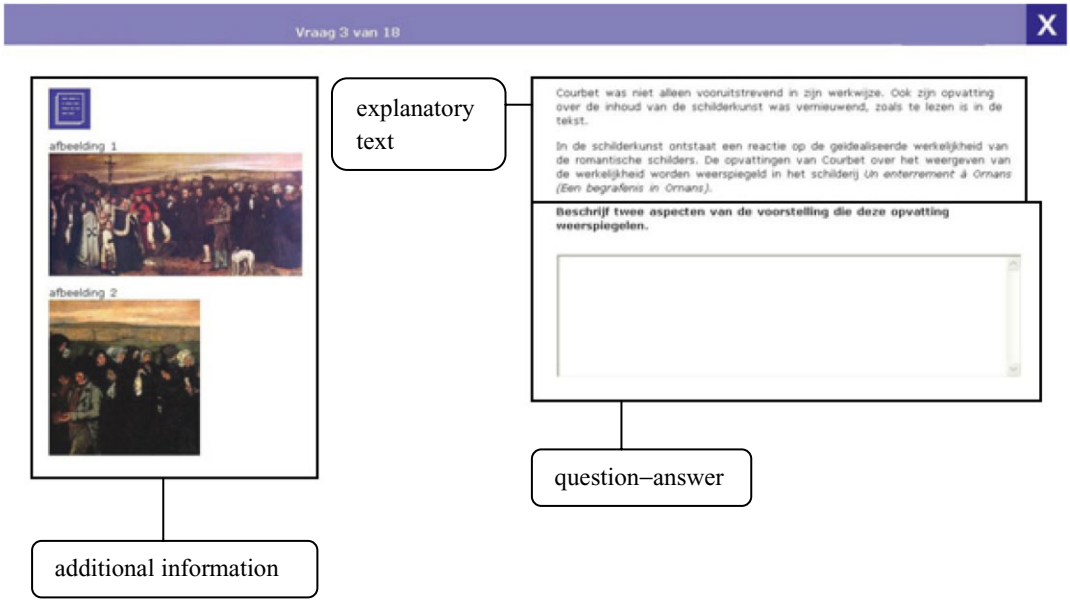

Figure 3: Division of the screen into three types of areas of interest, for which eye tracking parameters were summarized: explanatory text, additional information, and question–answer area

Results

All means and standard deviations are presented in Table 1. These descriptive data show that students solved about 50% of all tasks correctly in about 3.5 minutes per task, indicating a mean task difficulty. Moreover, students indicated to perceive the amount of mental effort (5) on a subjective rating scale (from 1 to 9).

Efficiency

A repeated-measures MANOVA was calculated with the within-subject factor 'presentation format' and the dependent variable 'efficiency per item'. Results show a main effect for presentation format, $F(5, 17) = 26.86$, $p < .01$. Univariate tests show that the split format always led to significantly more efficient results than the integrated format, that is for efficiency based on: subjective mental effort ratings ($efficiency_{split} = 0.00\ (1.18)$, $efficiency_{integrated} = -2.10\ (0.95)$, $F(1, 21) = 79.91$, $p < .01$), pupil dilation ($efficiency_{split} = 0.00\ (1.17)$, $efficiency_{integrate} = -2.10\ (0.88)$, $F(1, 21) = 122.06$, $p < .01$), fixation durations ($efficiency_{split} = 0.00\ (0.96)$, $efficiency_{integrated} = -2.10\ (0.77)$, $F(1, 21) = 137.00$, $p < .01$), silent pauses ($efficiency_{split} = 0.00\ (1.26)$, $efficiency_{integrated} = -2.10\ (0.96)$, $F(1, 21) = 55.09$, $p < .01$), and explicit utterances ($efficiency_{split} = 0.00\ (1.07)$, $efficiency_{integrated} = -2.10\ (0.84)$, $F(1, 21) = 65.88$, $p < .01$).

Mental effort measures

Spearman's correlations were calculated to investigate the relation between the different mental effort measures. Results show that mental effort indicated by fixation durations significantly

Table 1: Means and standard variations of all dependent variables

	Split format	*Integrated format*
Performance		
Percentage correct	50.38 (15.74)	43.75 (12.11)
Time on task in minutes	3.55 (1.28)	3.64 (0.94)
Mental effort		
Subjective rating[a]	5.01 (0.62)	5.01 (0.64)
Pupil dilation in mm	3.96 (0.45)	3.93 (0.43)
Fixation durations in milliseconds	299.43 (56.65)	282.52 (49.85)
Silent pauses in occurrences	4.07 (2.46)	4.16 (2.04)
Utterances of difficulties	0.36 (0.50)	0.28 (0.25)
Visual search		
Transitions	69.50 (30.36)	74.10 (17.84)
Time spent on AOIs in sec		
Explanatory text	23.70 (11.95)	18.72 (8.13)
Additional text info	32.06 (22.01)	44.28 (21.81)
Additional pictorial info	42.83 (22.10)	47.07 (22.24)
Question–answer area	29.55 (11.02)	25.38 (11.98)
Number of fixations on AOIs		
Explanatory text	87.22 (42.24)	71.93 (33.48)
Additional text info	132.06 (86.25)	179.64 (78.95)
Additional pictorial info	158.07 (83.70)	143.59 (67.17)
Question–answer area	92.49 (38.03)	91.10 (35.71)

[a]Rated on a scale ranging from 1 (very, very low effort) to 9 (very, very high effort).

correlates with effort indicated by pupil dilation ($r_s = -.42$, $p = .03$) and marginally with effort indicated by silent pauses ($r_s = -.34$, $p = .06$). Mental effort as indicated by subjective mental effort ratings marginally correlates with effort as indicated by explicit utterances ($r_s = .34$, $p = .06$) and effort indicated by silent pauses ($r_s = .30$, $p = .09$).

Visual search

Transitions

A one-way repeated measures ANOVA with the within-subject factor 'presentation format' and the dependent variable 'number of transitions per item' was calculated. The result showed no significant differences between formats, $F < 1$.

Time spent on AOIs

A repeated-measures MANOVA with the within-subject factor 'presentation format' and the dependent variable 'total fixation time spent on AOIs per item' was calculated. Results show a main effect for the presentation format, $F(4, 18) = 5.45, p < .01$. Univariate tests show that both formats did not lead to different viewing times on the 'explanatory text' AOI, $F(1, 21) = 2.16, p = .16$, or on the 'additional visualization' AOI, $F < 1$. Rather, when completing an item in integrated format, participants looked significantly longer at the 'additional text' AOI, $F(1, 21) = 6.84$, $p = .02$, and marginally shorter on the 'question–answer' AOI, $F(1, 21) = 4.06$, $p = .06$, as compared to completing an item in the split format.

Post hoc analyses

After obtaining these unexpected results, the actual content of the textual and pictorial information elements was analyzed. Each additional information element was coded in terms of whether it contained information that would lead to a higher test performance score when mentioned in the test answer. Results showed that seven of the eight items provided answer-relevant additional pictorial information elements. One item provided an illustrative picture that was not necessary to

Table 2: Relevance of the information provided in the additional textual information elements

	Amount of information (in words)	
Test item	*Relevant*	*Irrelevant*
1	110	80
2	—	196
3	—	91
4	—	302
5	84	63
6	—	147
7	—	89
8	—	338

complete the item. The additional textual information elements, on the other hand, were mostly unnecessary for completing the items. For only two of the eight items, part of the textual information was answer-relevant. In these items, part of the textual information needed to be combined with the pictorial information to complete the item. For example, in one item (see Figures 1 and 2) the societal view of the painter, described in the text, needed to be related to what was shown in his painting. All other items presented answer-irrelevant information (ie, mostly background information). For example, in one test item the life of a composer was described while the question was not about his life but about musical aspects of a certain piece that he had written. Table 2 shows the amount of relevant and irrelevant information in the additional textual information elements.

Discussion

The present study investigated the influence of the design of multimedia CBT environments on cognitive test load. Two leading theories on *learning*, namely CLT (Sweller *et al*, 1998) and the CTML (Mayer, 2005b) recommend taking the limited capacity of the human cognitive system into account when designing multimedia environments. One of their guidelines is to avoid splitting the learner's visual attention, by presenting related information spatially and temporally in close proximity to each other (split attention effect: Chandler & Sweller, 1991; Moreno & Mayer, 1999). Such an integrated presentation format is assumed to reduce visual search of relevant information and thus free up cognitive capacity as compared to a split format, which in turn results in better performance. The present study applied this principle to the design of multimedia tests. Students in an authentic, standardized national *Art Appreciation* exam were asked to answer testing items that were either presented in the original split format or in a new integrated format. Additionally, participants' visual search was captured by diverse eye tracking parameters; and mental effort was captured by means of subjective, physiological, and concurrent speech measures.

Results showed—contrary to expectations—that students performed more efficiently on test items presented in a split format than on items presented in an integrated format. This finding was identical independent of what type of mental effort measure was used to calculate efficiency. Analyses of the eye tracking parameters that captured the nature of visual search, helped unravel these unexpected findings. Participants did not differ in the way they looked at the explanatory text or the additional pictorial information, but they did look differently at the additional textual information and the question–answer area depending on the presentation format of the item. Specifically, participants spent more time looking at additional textual information when a test item was presented in an integrated format than in a split format. In contrast, when a testing item was presented in a split format participants tended to spent more time looking at the 'question–answer' area instead.

Analysis of the additional text information revealed that most of it was irrelevant to answering the test question. Hence, changing the presentation format did change participants' visual behavior in that they attended to all given information when it was presented in the new integrated format, while ignoring it when it was presented in the original split format.

In learning, irrelevant information is known to hamper performance (cf. negative effect of seductive details on learning: Abercrombie, 2013; or coherence principle: Mayer, 2005a). In testing irrelevant information may also hamper performance, however, on purpose. One goal of an exam might be to test one's ability to deal with large amounts of information of unknown relevance to their current question or of their ability to discern between main issues and side-issues. In the current Arts exam, this would mean that the student must decide whether supplementary information, such as biographical information about an artist or one of his/her paintings, would help answer the question about the main characteristics of the artist's style. This challenge is often present in real-world tasks. Marine zoologists, for example, must be able to distinguish between relevant and irrelevant features when classifying fish (Jarodzka, Scheiter, Gerjets & Van Gog, 2010). If this was the aim of this exam, the integrated format in the present study would have provided a more true testing result as participants did consider the additional information. In the split format, participants did not attend to the additional information and thus could not evaluate whether it was relevant or not. In other words, before making generalizations based on results of this study, one should carefully consider the aim of testing in this domain.

Another unexpected result was that participants did not differ in the amount of visual search—as indicated by transitions between different elements on the screen—executed on the two different item presentation formats. The prediction was that an integrated format would result in less need for visual search and thus lead to fewer transitions between different screen elements in comparison to a split format. The results suggest that—at least in testing—the assumption that a specific presentation format directly leads to a specific amount of visual search may be too simplistic. It is likely that participants tried to make sense of the additional information provided in the context of the testing item, which in turn led to more transitions than would have been necessary if the information was relevant to the task. Thus, the effect of presentation format on the amount of visual search required may be moderated by the thematic relevance of the provided information. Another explanation may be that participants the integrated format did not need to execute much visual search since all information was presented where needed. In the split format, participants behaved in diverse ways (as can be seen by the large standard deviations): some of them tried to integrate the information and executed much visual search, while others did not need to execute any visual search as they simply ignored the additional information (cf. Holsanova *et al*, 2008). Consequently, both formats may have led to little mean visual search, but for different reasons. Future research should investigate the reasons behind the visual search differences more in depth, for instance by applying a post hoc interview that is cued by specific eye movement behavior—or lack of it (cued retrospective recall; cf. Van Gog *et al*, 2005). In sum, we can conclude that more research is needed on multimedia CBT that should eventually result in a *cognitive theory of multimedia testing*, which in turn would deliver design guidelines for multimedia CBT.

Finally, a correlation between the different types of mental effort measures revealed interesting findings. Indirect measures such as the eye tracking parameters fixation duration and pupil dilation were related to the rather passive measure of silent pauses, while concrete utterances of mental overload were related to self-reports (and only weakly related to silent pauses). These findings may indicate that there is an explicit (indicated by self-reports and utterances) and an implicit (indicated by eye tracking parameters and silent pauses) aspect to mental effort. This assumption should be tested more thoroughly in future research with more indicators of mental effort.

A limitation of our study is that only one participant was male, further research should replicate these findings with an equal sex distribution to rule out possible sex effects.

Conclusions for theory and educational practice

From a technical perspective there are no obstacles to using multimedia in CBT. But from a pedagogical perspective, there are still many open questions as to how such multimedia CBT should be designed and used. The results of the present study reveal that the design of CBT environments making use of multimedia influences visual attention of students. In an integrated format, in contrast to a split format, people can be coerced to inspect all presented information. Even more important, the design affects the efficiency with which testees can complete the exams. However, this relation is not as simple as initially assumed. It turned out that the *content* of the presented information plays a crucial role, too. Based on this study it can be concluded that this content moderates the efficiency of CBT designs. If additional information is irrelevant, a split format where this information is ignored is more efficient as it helps testees to ignore it, while if it is relevant, an integrated format may be more efficient as it would help testees to integrate it. Further research should investigate this possible mediation. Nevertheless, in the split format students' attention was guided away from the additional information; it was easy for them to ignore or oversee this information. In the integrated format, however, students' attention was guided towards this additional information; they could make a deliberate decision whether to ignore it or not. Therefore, the integrated format is more favorable in our case. In the future, further guidelines for the use of multimedia in testing should be developed.

A related issue is the aim of the testing: is it important to provide participants with many different information elements of varying relevance to find the relevant ones or only to provide them with relevant information elements that they merely have to integrate? In the former case, the aim of a test is to see if testees can filter the relevant information from much irrelevant information by themselves. This is often necessary in professional education, such as controlling air traffic or diagnosis of medical images. In such professions, experts (ie, the air traffic controller, the diagnostician) must be able to discern/detect relevant information amongst many different irrelevant elements and then correctly interpret it (Balslev *et al*, 2012; eg, in medicine: Jaarsma, Jarodzka, Nap, Van Merriënboer, & Boshuizen, 2014; in air traffic control: Van Meeuwen *et al*, 2014). Hence, the testing situation must reflect this challenging real life scenario.

In the latter case, the aim is to test for acquired knowledge and skills of a student as directly as possible. This is often the case at school (or at the beginning of an educational trajectory) where the skill and knowledge level of students is still rather low. Hence, it is crucial to present multimedia and all other information in an efficient manner and avoid unnecessary cognitive processes, such as unnecessary search for the relevant information that may overwhelm the novice / student. Otherwise the assessor will not know why a student might score badly: due to a lack of skills or due to ineffective implementation of multimedia. Thus, when designing a multimedia test, one has also always to consider the aim of the test.

Based on the current study, there is no simple relation between the design of a CBT and testing efficiency. We can conclude that an integrated format makes students attend to all given information. In our case, however, as this information was not relevant for the given task, it hampered test performance. Thus, other factors such as the content of the presented information and the aims of the testing also have to be carefully considered.

Acknowledgement

The authors would like to thank Joke Hofstee (CITO) for her advice and for providing the testing material.

References

Abercrombie, S. (2013). Transfer effects of adding seductive details to case-based instruction. *Contemporary Educational Psychology*, *38*, 149–157. doi: 10.1016/j.cedpsych.2013.01.002.

Atkinson, R. C. & Shiffrin, R. M. (1968). Human memory: a proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds), *The psychology of learning and motivation: advances in research and theory* Vol. 2 (pp 89–192). New York: Academic Press.

Baddeley, A. D. (1992). Working memory. *Science*, *255*, 556–559. doi: 10.1126/science.1736359.

Balslev, T., Jarodzka, H., Holmqvist, K., De Grave, W. S., Muijtjens, A., Eika, B. *et al* (2012). Visual expertise in paediatric neurology. *European Journal of Paediatric Neurology*, *16*, 2, 161–166. doi: 10.1016/j.ejpn.2011.07.004.

Cerpa, N., Chandler, P. & Sweller, J. (1996). Some conditions under which integrated computer-based training software can facilitate learning. *Journal of Educational Computing Research*, *15*, 345–367. doi: 10.2190/MG7X-4J8N-CKYR-P06T.

Chandler, P. & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and instruction*, *8*, 293–332. doi: 10.1207/s1532690xci0804_2.

College voor Examens (CEVO). (2010). Syllabus 2010 kunst, havo. Retrieved May 16, 2014, from http://www.examenblad.nl/examenstof/syllabus-2010-kunst-algemeen-havo/2010/havo/f=/kunst_algemeen_havo_2010.pdf.

De Boer, N. (2009). *De computer bij de centrale examens. Duidelijk digitaal 2 [The computer at the national exams. Clearly digital 2]*. Retrieved September 19, 2013, from http://www.cito.nl/VO/ce/compex/introductie/cve_comp_bij_ce_duidelijk_digitaal_2.pdf.

De Jong, T. (2010). Cognitive load theory, educational research, and instructional design: some food for thought. *Instructional Science*, *38*, 105–134.

Ericsson, K. A. & Simon, H. A. (1993). *Protocol analysis: verbal reports as data*. Cambridge, MA: MIT Press.

Hannus, M. & Hyönä, J. (1999). Utilization of illustrations during learning of science textbook passages among low- and high-ability children. *Contemporary Educational Psychology*, *24*, 95–123. doi: 10.1006/ceps.1998.0987.

Hegarty, M. & Just, M. A. (1993). Constructing mental models of machines from text and diagrams. *Journal of Memory and Language*, *32*, 717–717.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H. & Van de Weijer, J. (2011). *Eye tracking: a comprehensive guide to methods and measures*. Oxford, UK: Oxford University Press.

Holsanova, J., Holmberg, N. & Holmqvist, K. (2008). Reading information graphics: the role of spatial contiguity and dual attentional guidance. *Applied Cognitive Psychology*, *23*, 1215–1226. doi: 10.1002/acp.1525.

Jaarsma, T., Jarodzka, H., Nap, M., Van Merriënboer, J. J. G. & Boshuizen, H. P. A. (2014). Expertise differences under the microscope: processing histopathological slides. *Medical Education*, *48*, 3, 292–300. doi: 10.1111/medu.12385.

Jarodzka, H., Scheiter, K., Gerjets, P. & Van Gog, T. (2010). In the eyes of the beholder: how experts and novices interpret dynamic stimuli. *Journal of Learning and Instruction*, *20*, 146–154. doi: 10.1016/j.learninstruc.2009.02.019.

Just, M. & Carpenter, P. (1976). Eye fixations and cognitive processes. *Cognitive Psychology*, *8*, 441–480. doi: 10.1016/0010-0285(76)90015-3.

Kalyuga, S., Chandler, P. & Sweller, J. (1999). Managing split attention and redundancy in multimedia instruction. *Applied Cognitive Psychology*, *13*, 351–371. doi: 10.1002/(SICI)1099-0720(199908)13:4<351::AID-ACP589>3.0.CO;2-6.

Klingner, J., Tversky, B. & Hanrahan, P. (2011). Effects of visual and verbal presentation on cognitive load in vigilance, memory, and arithmetic tasks. *Psychophysiology*, *48*, 323–332. doi: 10.1111/j.1469-8986.2010.01069.x.

Loftus, G. F. (1981). Tachistoscopic simulations of eye fixations on pictures. *Journal of Experimental Psychology. Human Learning and Memory*, *7*, 369–376. doi: 10.1037//0278-7393.7.5.369.

Mayer, R. E. (2001). *Multimedia learning*. New York: Cambridge University Press.

Mayer, R. E. (2005a). *The Cambridge handbook of multimedia learning*. New York: Cambridge University Press.

Mayer, R. E. (2005b). Cognitive theory of multimedia learning. In R. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp 31–48). New York: Cambridge University Press.

Miller, G. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, *63*, 81–97.

Moreno, R. & Mayer, R. E. (1999). Cognitive principles of multimedia learning: the role of modality and contiguity. *Journal of Educational Psychology*, *91*, 358–368. doi: 10.1037//0022-0663.91.2.358.

Müller, C., Grossmann-Hutter, B., Jameson, A., Rummer, R. & Wittig, F. (2001). Recognizing time pressure and cognitive load on the basis of speech: an experimental study. In M. Bauer, P. J. Gmytrasiewicz & J. Vassileva (Eds), *Proceedings of the 8th international conference on user modeling 2001* (pp 24–33). London, UK: Springer-Verlag.

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. doi: 10.1037//0022-0663.84.4.429.

Parshall, C. G., Spray, J. A., Kalohn, J. C. & Davey, T. (2002). *Practical considerations in computer-based testing*. New York: Springer Verlag.

Parshall, C. G., Harmes, J. C., Davey, T. & Pashley, P. J. (2010). Innovative items for computerized testing. In W. J. van der Linden & C. A. W. Glas (Eds), *Elements of adaptive testing* (pp 215–230). New York: Springer.

Sweller, J., Van Merriënboer, J. J. G. & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychological Review, 10*, 251–296. doi: 10.1023/B:TRUC.0000021808.72598.4d.

Tobii 1750 [Apparatus]. (2003). Danderyd, Sweden: Tobii Technology AB.

Tobii Studio [Software]. (2007). Danderyd, Sweden: Tobii Technology AB.

Tuovinen, J. E. & Paas, F. (2004). Exploring multidimensional approaches to the efficiency of instructional conditions. *Instructional Science, 32*, 133–152.

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. doi: 10.1080/02724980343000189.

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. doi: 10.1037/1076-898X.11.4.237.

Van Meeuwen, L. W., Jarodzka, H., Brand-Gruwel, S., Kirschner, P. A., De Bock, J. J. P. R. & Van Merriënboer, J. J. G. (2014). Identification of effective visual problem solving strategies in a complex visual domain. *Learning and Instruction, 32*, 10–21. doi: 10.1016/j.learninstruc.2014.01.004.

Van Mierlo, C. M., Jarodzka, H., Kirschner, F. & Kirschner, P. A. (2012). Cognitive load theory and e-learning. In Z. Yan (Ed.), *Encyclopedia of cyber behavior* (pp 1178–1211). Hershey, PA: IGI Global.

Van Orden, K. F., Limbert, W., Makeig, S. & Jung, T. P. (2001). Eye activity correlates of workload during a visuospatial memory task. *Human Factors, 43*, 1, 111–121.

Ward, M. & Sweller, J. (1990). Structuring effective worked examples. *Cognition and instruction, 7*, 1–39. doi: 10.1207/s1532690xci0701_1.

Yin, B. & Chen, F. (2007). Towards automatic cognitive load measurement from speech analysis. In J. A. Jacko (Ed.), *Proceedings of the 12th international conference on Human-computer interaction: interaction design and usability* (pp 1011–1020). Berlin, Heidelberg: Springer-Verlag.