Computers in Human Behavior 62 (2016) 394–403

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Full length article

Medical students' cognitive load in volumetric image interpretation: Insights from human-computer interaction and eye movements

Bobby G. Stuijfzand [a], [*], Marieke F. van der Schaaf [a], Femke C. Kirschner [a], Cécile J. Ravesloot [b], Anouk van der Gijp [b], Koen L. Vincken [c]

[a] Department of Education, Utrecht University, Heidelberglaan 1, 3584 CS Utrecht, The Netherlands
[b] Department of Radiology, University Medical Centre Utrecht, Heidelberglaan 100, 3584 CX Utrecht, The Netherlands
[c] Image Sciences Institute, University Medical Centre Utrecht, Heidelberglaan 100, 3584 CX Utrecht, The Netherlands

ARTICLE INFO

Article history:
Received 16 December 2015
Received in revised form
1 April 2016
Accepted 5 April 2016
Available online 13 April 2016

Keywords:
Volumetric image interpretation
Human-computer interaction
Eye tracking
Cognitive load
Medical education

ABSTRACT

Medical image interpretation is moving from using 2D- to volumetric images, thereby changing the cognitive and perceptual processes involved. This is expected to affect medical students' experienced cognitive load, while learning image interpretation skills. With two studies this explorative research investigated whether measures inherent to image interpretation, i.e. human-computer interaction and eye tracking, relate to cognitive load. Subsequently, it investigated effects of volumetric image interpretation on second-year medical students' cognitive load. Study 1 measured human-computer interactions of participants during two volumetric image interpretation tasks. Using structural equation modelling, the latent variable ‘volumetric image information’ was identified from the data, which significantly predicted self-reported mental effort as a measure of cognitive load. Study 2 measured participants' eye movements during multiple 2D and volumetric image interpretation tasks. Multilevel analysis showed that time to locate a relevant structure in an image was significantly related to pupil dilation, as a proxy for cognitive load. It is discussed how combining human-computer interaction and eye tracking allows for comprehensive measurement of cognitive load. Combining such measures in a single model would allow for disentangling unique sources of cognitive load, leading to recommendations for implementation of volumetric image interpretation in the medical education curriculum.

1. Introduction

Over the past two decades, cross-sectional image interpretation in medicine has shifted from using 2D images to volumetric images to diagnose patients. A volumetric image involves a volumetric medical scan, e.g., computed tomography [CT] or magnetic resonance imaging [MRI] that can be sliced up in many cross sections (i.e., ‘slices’) forming a stack of images. The user can scroll through a volumetric image from various angles and in various contrast settings, creating a 3-dimensional representation of the scanned structure. This shift has changed the task of medical image interpretation. A tiled set of 2D-images is static and contains less information than a volumetric image (Krupinski, 2011; Krupinski et al., 2012). Interpretation of volumetric images is more dynamic, involving an increase in both visual information processing and human-computer interaction (HCI, Andriole et al., 2011; Krupinski, 2010; Reiner, Siegel, & Siddiqui, 2003). Expert skill in image interpretation, including volumetric image interpretation, is crucial to avoid medical diagnostic errors (Donald & Barnard, 2012; Pinto et al., 2011), and thus volumetric images are now increasingly being used in medical education as well (Ravesloot, van der Gijp, et al., 2015; Rengier et al., 2013; van der Gijp et al., 2015).

Recent research in medical education highlights the effects this shift has had on students engaged in the image interpretation task. Radiology clerks take more time, and engage in more and different cognitive processes when interpreting volumetric images than 2D images (van der Gijp et al., 2015). Medical students report volumetric images to be more representative of clinical practice and perceive them to be easier to interpret than their 2D-counterparts.

[*] Corresponding author. Present address: School of Experimental Psychology, University of Bristol, 12a Priory Road, Bristol, BS8 1TU, United Kingdom.

E-mail addresses: bg.stuijfzand@bristol.ac.uk (B.G. Stuijfzand), m.f.vanderschaaf@uu.nl (M.F. van der Schaaf), f.c.kirschner@uu.nl (F.C. Kirschner), c.j.ravesloot@umcutrecht.nl (C.J. Ravesloot), a.vandergijp@umcutrecht.nl (A. van der Gijp), k.vincken@umcutrecht.nl (K.L. Vincken).

http://dx.doi.org/10.1016/j.chb.2016.04.015
0747-5632/© 2016 Elsevier Ltd. All rights reserved.

Interestingly though, performance of these students on interpretation of volumetric images was lower than on 2D images (Ravesloot, van der Gijp, et al., 2015; Ravesloot, van der Schaaf, et al., 2015).

A little studied aspect that may be particularly affected by volumetric image interpretation is students' cognitive load. Cognitive load, i.e., demand on human working memory, plays a pivotal part in the construction, elaboration, and automation of knowledge structures (i.e., schemas; Chi, Glaser, & Rees, 1982) in long-term memory (Sweller, van Merriënboer, & Paas, 1998; van Merriënboer & Sweller, 2005). Human's working memory capacity is limited and the cognitive load experienced is directly influenced by the information that the student needs to process and the schemas the student already possesses. A skilled student is better able to ignore task-irrelevant information and integrate new information with existing schemas and will therefore experience less cognitive load in a complex task than an unskilled student. Although cognitive load as a result from engagement in a learning task can be beneficial as it involves processing of task-relevant information, cognitive overload has shown to be detrimental for learning performance (Sweller, 2004; for an elaborate background on cognitive load theory in medical education, see van Merriënboer & Sweller, 2010). Previous research has identified relationships between cognitive load and visual information, simulated 3D environments, and human computer interaction in digital learning environments (e.g., Hollender, Hofmann, Deneke, & Schmitz, 2010; Mayer & Moreno, 2003; Ruiz, Taib, & Chen, 2011; Ruiz, Taib, Shi, Choi, & Chen, 2007; van der Land, Schouten, Feldberg, van den Hooff, & Huysman, 2013); however, to our knowledge little research is available in the context of medical image interpretation.

The present paper aims to shed light onto how volumetric image interpretation affects cognitive load experienced by medical students. Measures that indicate visual information processing and human-computer interaction are combined, and their common variance is used to predict cognitive load measures to: (1) investigate whether these measures can be utilised as indirect objective measures of cognitive load and, (2) to investigate how volumetric image interpretation by medical students affects their cognitive load.

1.1. Image interpretation in medical education

Medical image interpretation involves detecting and interpreting abnormalities in images of the human body for diagnostic purposes (Krupinski, 2010; Norman, Coblentz, Brooks, & Babcook, 1992; Taylor, 2007). Traditionally, assessment of students' image interpretation skills often involved interpreting single 2D images. In volumetric images, students do not examine one image to find a relevant structure, but must view a whole stack of slices, use an appropriate contrast setting, and in some cases adjust the angle to identify a structure. As a consequence they have to examine more information, inherently make more considerations regarding the relevancy of this information, while manipulating the image (Krupinski, 2010; van der Gijp et al., 2015). During image interpretation students have to cognitively link all the slices together in order to create a mental 3D representation of the body, which requires spatial skills and cognitive capacity of students (Krupinski, 2010; Stull, Hegarty, & Mayer, 2009). This increase in visual information and human-computer interaction when using volumetric images has been related to an increase in cognitive load in other contexts (van Merriënboer & Sweller, 2005).

Conversely, volumetric image interpretation may also decrease cognitive load. The possibility of examining the anatomical structure and its relative position from multiple angles can arguably provide the student with additional contextual information, i.e. the student does not need to infer the shape, size and position of a structure based on one 2D image (Ellis et al., 2006; Hegarty, Keehner, Cohen, Montello, & Lippa, 2007; van der Land et al., 2013). This contextual information allows for less specific prior knowledge needed for image comprehension (van Merriënboer & Sweller, 2010). As a result, it is currently unclear how volumetric image interpretation would affect cognitive load.

1.2. Measuring cognitive load in image interpretation

A wide variety of measures are utilised for measuring cognitive load, such as dual-task methodology (Brünken, Steinbacher, Schnotz, Plass, & Leutner, 2002), physiological measures (Antonenko, Paas, Grabner, & van Gog, 2010; DeLeeuw & Mayer, 2008; Nourbakhsh, Wang, & Chen, 2013), and self-report ratings (Kirschner, Paas, & Kirschner, 2009). However, these measures only provide a quantitative indication of cognitive load (Sweller, Ayres, & Kalyuga, 2011) but are uninformative of what causes this cognitive load. Using indirect objective measures that are specific to the (volumetric) image interpretation task, and relate these to validated subjective and physiological measures for cognitive load (e.g., DeLeeuw & Mayer, 2008) may address this (Martin, 2014). Indirect objective measures are direct reflections of task behaviour that bear a relationship with cognitive load, but this relationship may be mediated or moderated by other variables such as skill or task-performance (Brünken, Plass, & Leutner, 2003). If common variance of image interpretation task-specific objective measures has a relationship with validated measures of cognitive load while taking into account mediators and moderators, this would support using these measures for disentangling cognitive load in image interpretation. The nature of each of the contributing measures can then highlight what aspects of the task are related to cognitive load.

1.3. Approach

In the first study, variables are calculated from recorded human-computer interaction of participants engaged in volumetric image interpretation tasks. Logging participants' interactions within the learning environment reveals how many slices are displayed due to scrolling through the image, how many viewing angle changes are made, how long it takes to locate the relevant slices, how much time is spent on relevant vs. irrelevant slices, and how long a student takes to finish a task (Vincken & Ravesloot, 2010). Although time to finish a task has been previously related to various types of cognitive load in other contexts, the usage of this data in this context is new and effects on cognitive load are unknown (Brünken et al., 2003). As human-computer interaction variables are only conceptualised as indirect objective measures of cognitive load, other potential factors in these relationships must be considered. For example, it should be acknowledged that experts are better in deciding which information is relevant than novices (Eva, Norman, Neville, Wood, & Brooks, 2002; Lesgold et al., 1988; van Gog, Kester, Nievelstein, Giesbers, & Paas, 2009), and are quicker to find abnormalities in medical images (Kok et al., 2015). Although the medical students participating in the current study are all at similar stages of their training, performance differences caused by differential skill development are likely. As a result, there is a potential influence of performance on the relationship of information exposure and cognitive load (Brünken et al., 2003). The first study therefore includes a measure of test-performance in image interpretation to control for this and to investigate a potential moderation in the relationship between human-computer interaction and cognitive load.

In a second study, eye tracking is used to provide more in-depth information on what a medical student examines while interpreting a medical image and how this relates to cognitive load. In previous studies, participants who took longer to locate relevant areas in images experienced higher task-complexity, which is associated with higher cognitive load (Corbetta & Shulman, 2002; De Fockert, Rees, Frith, & Lavie, 2001). The human-computer interaction and eye tracking measures are related to two well-established measures of cognitive load; the self-reporting mental effort scale and pupil-dilation (e.g., Bailey & Iqbal, 2008; Paas, 1992; Palinko, Kun, Shyrokov, & Heeman, 2010; Zheng & Cook, 2012).

1.4. Hypotheses

To investigate how volumetric image interpretation affects medical students' cognitive load using human-computer interaction and visual information measures, Study 1 uses human-computer interaction data to measure participants' behaviour in a volumetric image interpretation task and investigates the following hypotheses:

H1. Variables indicating human-computer interaction have common variance, resulting in the latent variable volumetric image information.

H2. The latent variable volumetric image information predicts self-reported mental effort as a measure of cognitive load.

H3. Overall test performance has a negative relationship with self-reported mental effort.

H4. Overall test performance has an interaction effect with volumetric image information on mental effort.

In Study 2, eye tracking is employed in a new sample, to investigate more specifically the relationship between the nature of the examined visual information and cognitive load:

H5. The time to locate a relevant structure is positively related to self-reported mental effort and pupil dilation, as measures of cognitive load.

2. Method Study 1

2.1. Sample

The sample consisted of second-year medical students taking a radiology test, which was one of the components of an end of second year practical examination. The complete test-group consisted of 92 students, of which 67 approved of usage of their data for scientific purposes. The resulting sample ($N = 67$) had age ranging from 19.30 to 56.19 years ($M = 22.57$, $SD = 5.14$) and included 40 females. Twenty-two students were new to the test; the remaining 45 took a similar test three months earlier. Reasons for retaking the test: (1) One student required a retake due to failing the previous radiology test, (2) 44 students required a retake due to failing another part of the practical examination. An independent samples t-test (equal variances not assumed) on the current test scores showed that the two subsamples ($n_1 = 22$, $n_2 = 45$) did not significantly differ in their achievements: $\Delta M = -0.12$, $t(69.82) = -0.30$, $p = 0.77$ and thus repeating the test did not influence the homogeneity of sample performance.

2.2. Materials

2.2.1. Radiology test

The radiology test consisted of 40 items, of which 20 were 2D image interpretation items, and 20 were volumetric image interpretation items. Both could contain a question of one of the following types: (1) locate and mark a structure mentioned in the question, or (2) identify a highlighted structure.

2.2.2. Assessment tool

The test was digitally administered using VQuest (Vincken & Ravesloot, 2010). VQuest is a software package developed at University Medical Centre Utrecht that can display 2D and volumetric images. VQuest runs in full-screen mode and the medical image covers 64% of the screen. The image is scaled, meaning there is no need to scroll in the horizontal and vertical directions. To navigate through different slices of the volumetric image the student can either: use the scroll wheel on a mouse, move a mouse while pressing the left mouse-button, or use the Page Up/Down-keys on a keyboard. Left of the image buttons are available to change the viewing angle (i.e., axial, sagittal, or coronal), and underneath that questions and answer possibilities are displayed. Above the image, a percentage is displayed, indicating the student's progression through the test. Also above the image a button is available through which a student can access different items within the test.

2.2.3. Task-description

The item used for main analysis in this study (Task 1) was the final item in the test and concerned the volumetric image of a CT-scan of a human's abdomen. Students were asked to locate the portal vein, which is a blood vessel from the intestines to the liver. To complete the item students had to navigate through the volumetric image and find a slice with the portal vein present. When located, students had to place a digital marker in the portal vein. The marker could be replaced at any time during the whole test, with only the last marker being scored. The task score, which contributed to the overall test performance, was dichotomous: the marker was placed correctly or not. Directly after final marker placement, students indicated the mental effort required to complete this specific image interpretation task. Task 1 was of reasonable difficulty with a p-value of $p = 0.72$, (i.e., proportion of students answering the task correctly) and sufficiently discriminating with an item-total correlation of $r = 0.24$ (Evers, Lucassen, Meijer, & Sijtsma, 2010).

A second item provided data to validate results of the analysis (Task 2). The item's appearance in the test was randomised for each student, but always preceded Task 1. This task was similar to the primary task, and involved a CT-scan of the thorax. Students were asked to mark the right erector spinae muscle (i.e., long back muscle). It differed from task 1 as no mental effort rating was administered. Task 2 was of reasonable difficulty with a p-value of $p = 0.76$, and well discriminating with an item-total correlation of $r = 0.35$ (Evers et al., 2010).

2.3. Measures

2.3.1. Cognitive load

Cognitive load was operationalised using a self-reported one-item mental effort scale (Paas, 1992). The item was formulated in Dutch and slightly altered to fit the context of the study, and translates into ‘Please indicate, on the scale depicted below, how much effort it took you to complete the task?’ Students answered the item on a 9-point scale, ranging from (1) very very low mental effort to (9) very very high mental effort. The scale is a frequently used and reliable measure for cognitive load (Paas, Tuovinen, Tabbers, & van Gerven, 2003; Paas, van Merriënboer, & Adam, 1994) that has previously demonstrated discriminant validity (Sweller et al., 2011). The measure was chosen given its sensitivity and non-invasive nature in comparison with other measures of cognitive load (Sweller et al., 2011).

2.3.2. Human-computer interaction

VQuest logs activities performed within the above-described tasks. The information from these logs was used to construct the latent variable volumetric image information. The following variables were extracted from the logs per task: (1) Total time spent in the task, measured in seconds (min: 0, max: inf.). This was calculated by subtracting time taken to start scrolling from time of finishing the task to achieve that only information forthcoming from the volumetric nature of the task was included (i.e., time taken to read the question, and time necessary to load the images, was excluded). (2) Standardised amount of slices displayed; a calculated Z-score on the amount of slices that have been displayed due to scrolling through the volumetric image. Standardising was considered necessary as amount of task-required scrolling differed between tasks. (3) Number of angle changes (min: 0, max: inf.). (4) Proportion of time spent on relevant slices (min: 0, max: 1), with respect to total time in the task. A relevant slice was defined as one that displays the structure mentioned in the question. (5) Time taken to locate first relevant slice, in seconds (min: 0, max: inf.). This variable was measured from the moment a participant started scrolling.

2.3.3. Overall test performance

Overall test performance was operationalised by the obtained grade of a student on the radiology test (min: 1, max: 10). The test was sufficiently reliable with Cronbach's alpha = 0.74 (Evers et al., 2010), indicating that test-performance was a consistent measure for high- and low-test performers.

2.4. Procedure

2.4.1. Data-collection

The data was collected in the summer of 2012 at Utrecht University, The Netherlands. Before data collection commenced, ethical approval was obtained from the Ethical Review Board of the Netherlands Association for Medical Education and all participants signed informed consent forms. Participation was voluntary and no compensation was offered.

2.4.2. Test-environment

Students were asked to read and sign the informed consent form a priori. During the test, instructors were present for technical assistance. Students had been able to practice with the assessment tool VQuest previously, and paper-instructions on how to use the VQuest programme were available for each student at the time of the test. Students had 90 min to complete the test.

2.5. Data analysis

2.5.1. Assumptions

Included variables were screened for missing values, presence of outliers and normality. Number of angle changes was not assumed to be normal, given that the usage of this functionality was optional, i.e. not every participant used this, hereby creating a bottom-effect. Due to this expected non-normality, in all analyses maximum likelihood estimation with robust standard errors was used, and chi-square difference tests were corrected following Satorra (2000). Interpretation of model fit indices, resulting from maximum likelihood estimation, followed Hu and Bentler (1999). Given the small sample size model fit indices should be interpreted with care. The CFI index is relatively independent from sample size, and SRMR is only positively biased with large sample sizes. Therefore these indices were given priority in assessing model fit (Byrne, 2012; Fan, Thompson, & Wang, 1999).

2.5.2. Hypothesis 1

To establish whether an underlying latent variable volumetric image information could be identified from human-computer interaction data, an exploratory factor analysis (EFA) was conducted on the following variables from Task 1: Total time spent in the task, standardised amount of slices displayed, number of angle changes, proportion of time spent on relevant slices, and time taken to locate first relevant slice. Items with a sufficient factor loading (i.e., >0.30) were included in the latent variable. The reliability of the latent variable was assessed using standardised Cronbach's alpha, due to differing variable scales. Subsequently, a confirmatory factor analysis (CFA) was conducted to establish whether one latent variable, emerging from the EFA, fitted the data of Task 1.

Next, Task 1 data was merged with Task 2 data to investigate measurement invariance. This assessed whether the latent variable could be found in different volumetric image interpretation tasks within the test, which would provide support for the validity of the latent variable. Measurement invariance is tenable when the relations between the variables and the latent variable are the same across tasks (Koh & Zumbo, 2008). To assess this, a chi-square difference test compared two models: (1) a model in which factor loadings were constrained to be the same between tasks, and (2) a model where factor loadings were allowed to vary between tasks. An insignificant chi-square would indicate the model does not deteriorate by adding constraints; hence, these constraints are permissible and measurement invariance is supported.

2.5.3. Hypothesis 2–3

Using Task 1 data, hypotheses 2 and 3 were tested. A structural equation model was specified with self-reported mental effort as dependent variable and the latent variable volumetric image information and overall test-performance as independent variables. The model was used to assess significance of relationships between volumetric image information and self-reported mental effort, as well as between overall test performance and self-reported mental effort.

2.5.4. Hypothesis 4

An interaction between volumetric image information and overall test performance was added to the previous model to test hypothesis 4. Significance of the interaction effect would provide evidence for differing effects of the latent variable volumetric image information on self-reported mental effort for students of varying levels of proficiency in image interpretation.

3. Results Study 1

3.1. Assumptions

No missing values were present. Two identified extreme cases showed long periods of non-activity in VQuest's log-files and were deleted from the sample as outliers. All but three variables were normally distributed. Square root transformations were applied to time in the task and time taken to locate first relevant slice to comply with normality. Identified non-normality of number of angle changes was not transformed since this was expected.

3.2. Descriptive statistics

Descriptive statistics for task 1 are presented in Table 1.

3.3. Hypothesis 1

3.3.1. Exploratory factor analysis

The EFA was conducted on the five human-computer interaction

Table 1
Descriptive statistics Study 1 for variables in task 1.

Variable	Min	Max	*M*	*SD*
Mental effort	1.00	9.00	5.26	1.59
Grade	3.30	9.70	6.18	1.28
Total time spent in the task	3.32	17.75	7.78	3.22
Standardised amount of slices displayed	−1.02	1.95	−0.09	0.81
Number of angle changes	0.00	8.00	1.20	2.16
Proportion of time spent on relevant slices	0.00	0.79	0.44	0.18
Time taken to locate first relevant slice	0.00	5.57	2.08	1.32

Note. $n = 65$.

variables. All variables correlated > 0.3 with at least one of the other variables, suggesting an underlying latent variable. Inspection of time to locate first relevant slice showed a negative residual variance and a factor loading of −1.65. Such figures indicate data-issues, leading to exclusion of this variable. Rerunning analysis with the remaining four variables resulted in a one-factor model with satisfactory model fit of $\chi^2(2) = 1.76$, $p = 0.42$, $RSMEA < 0.001$, $SRMR = 0.03$ and a factor determinacy of 0.94. On this factor, three out of four variables had factor loadings > 0.3, shown in Table 2. Proportion of time spent on relevant slices did not show a sufficient factor loading (−0.03), and thus was not considered a good indicator of the factor. This variable was not included in further analyses. Reliability analysis, using standardised Cronbach's alpha, was conducted on the latent variable consisting of three variables. As shown in Table 2 the latent variable was sufficiently reliable (Evers et al., 2010).

3.3.2. Confirmatory factor analysis

The CFA was conducted to test the measurement model of the latent variable from the EFA. All factor loadings were significant, and model fit was good: $\chi^2(1) = 0.84$, $p = 0.36$, $CFI = 1.00$, $TLI = 1.02$, $RSMEA < 0.001$, $SRMR = 0.03$. The good fit of the measurement model supports hypothesis one, that an underlying latent variable volumetric image information can be identified from the data.

3.3.3. Measurement invariance

Next, Task 2 data was added to the dataset to test for measurement invariance. This assessed whether the latent variable volumetric image information is task-dependent, or can be identified in different volumetric image interpretation tasks within this test. Good model fit was retained after inclusion of Task 2 data, as shown in Table 3. Second, adding constraints on factor loadings across Task 1 and Task 2 did not significantly deteriorate the model, as indicated by the insignificant chi-square difference (see Table 3). This shows that in both tasks, variables have the same relationship with volumetric image information, hereby supporting measurement invariance.

Table 2
Factor Loadings and Reliability of the latent variable.

	λ[a]	α[b] (standardised)
Factor		0.70
Total time spent in the task	0.84	
Standardised amount of slices displayed	0.91	
Number of angle changes	0.33	

Note. $n = 65$.
[a] λ = factor loadings.
[b] α = Cronbach's alpha.

Table 3
Model fit and chi-square statistics for the measurement invariance analysis.

Model	χ^2	*df*	*CFI*	*TLI*	*RSMEA*	*SRMR*	$\Delta\chi^2$
Free model (Task 1 & Task 2)	1.04	2	1.00	1.06	<0.001	0.02	
Constrained factor loadings	3.36	5	1.00	1.04	<0.001	0.08	2.32

Note. $n = 65$. All χ^2 are $p > 0.05$.

3.4. Hypothesis 2–3

3.4.1. Structural equation model

To test whether volumetric image information and overall test-performance predicted self-reported mental effort, a structural equation model was fitted on the Task 1 data. The structural model had good model fit: $\chi^2(6) = 3.61$, $p = 0.63$, $CFI = 1.00$, $TLI = 1.07$, $RSMEA < 0.001$, $SRMR = 0.04$. Within the model volumetric image information significantly predicted self-reported mental effort with $\beta = 0.40$, $SE = 0.18$, $p = 0.03$. There was no significant relation between test-performance and self-reported mental effort with $\beta = -0.03$, $SE = 0.02$, $p = 0.11$. The explained variance of self-reported mental effort was insignificant with $R^2 = 0.10$, $SE = 0.07$, $p = 0.16$.

3.5. Hypothesis 4

To test hypothesis 4, an interaction effect between volumetric image information and overall test-performance was added to the structural model. The interaction effect on self-reported mental effort was not significant ($B = 0.00$, $SE = 0.01$, $p = 0.61$) however inclusion of this effect in the model rendered the previously identified direct effect of volumetric image interpretation to self-reported mental effort insignificant ($B = 0.37$, $SE = 0.46$, $p = 0.43$). The direct effect of test-performance on self-reported mental remained insignificant ($B = -0.03$, $SE = 0.02$, $p = 0.06$).

4. Method Study 2

4.1. Sample

Participants were ten second year medical students. In the case of two students, data of the eye tracker could not be synchronised with VQuest logfile data due to calibration issues, reducing the sample to eight participants. Seven were female, and age ranged from 19 to 29 ($M = 22.20$, $SD = 3.52$). All participants had previously passed a radiology test required for their curriculum.

4.2. Materials

4.2.1. Radiology test

The tasks used for data-analysis were extracted from a similar radiology test as in Study 1. In contrast to Study 1 however, the test was administered only for scientific purposes, results were of no academic importance to the students.

4.2.2. Eye tracking instrument

For measuring students' eye-movements, a Tobii T60 was used. The device consists of a 24″ TFT screen, displaying a screen resolution of 1280 × 1024. The inbuilt eye tracker measures participants' eye movements at a rate of 60 Hz with accuracy of approximately 0.5°. This study utilised the software accompanying the instrument (Tobii Studio 3.2) for processing fixations and saccades (Tobii technology, 2010).

4.2.3. *Assessment tool*

For running the test, as in Study 1, VQuest (Vincken & Ravesloot, 2010) was used.

4.2.4. *Task-description*

Eight items in the test were available for data analysis. Four of the items were 2D image interpretation questions, and four of the items were volumetric image interpretation questions. In both types, students had to mark anatomic structures, similar to the task in Study 1. Of the 2D questions, task difficulty in *p*-values ranged from 0.22 to 0.99 and discrimination in item–total correlations ranged from −0.13 to 0.25. Of the volumetric questions *p*–values ranged from 0.36 to 0.94, and item-total correlations ranged from 0.22 to 0.48. Thus, difficulty between 2D and volumetric tasks was approximately the same though the discriminative value differed.

4.3. *Measures*

4.3.1. *Cognitive load*

As in Study 1, cognitive load was operationalised by the self-reported one-item mental effort scale (Paas, 1992). The self-reporting scale could become intrusive for the student if presented after all 40 items in the test and it was therefore decided to only administer the scale after eight selected items. The items appeared in random order during the test.

As a second measure on these eight items, average pupil dilation during a task was included as an indicator of cognitive load. Pupil dilation has shown to be a sensitive measure for cognitive load, although its sensitivity decreases with age (Paas et al., 2003; van Gerven, Paas, van Merriënboer, & Schmidt, 2004). Considering the young sample usage here was deemed appropriate.

4.3.2. *Time to locate relevant structure*

This variable was operationalised by the percentage of time in the task spent before locating the relevant structure. Locating a relevant structure is defined as the first gaze-coordinates falling within the coordinates of the structure in which the marker should be placed. Sometimes a participant might not locate the relevant structure; given that this is still considered relevant information for the hypothesis, in that case the total time in the task was used (i.e., a participant spent 100% of the time in the task without locating the relevant structure).

4.4. *Procedure*

4.4.1. *Data-collection*

The data was collected in the summer of 2012 at Utrecht University, The Netherlands.

The same procedures regarding ethical approval and informed consent were followed as in Study 1. Students were approached for voluntary participation by email and students were compensated for their participation with a voucher.

4.4.2. *Test-environment*

Students took the test individually in a closed, dimly lit room. Before the radiology test started, students signed an informed consent form and were instructed by a researcher how to use the VQuest software in combination with the eye tracker. They were then seated in a chair fixed in its position to the eye tracker. Participants' eyes were located approximately 80 cm from the screen, and participants were asked to keep their head as stable as possible in order to make optimal eye tracking possible. After this, the eye tracker was calibrated and subsequently the test commenced. During the test, only the student was present in the room, but a researcher was available in an adjacent room for technical assistance. There was no time limit, but no student exceeded the time limit in Study 1 (90 min).

4.5. *Data analysis*

4.5.1. *Assumptions*

Data was screened for missing values and outliers. Independency of the observations was violated due to inclusion of eight tasks in the analysis for each participant, which was addressed by conducting multilevel analyses for both dependent variables. Participants ($n = 8$) were specified as grouping variable (level 2). Self-reported mental effort per task ($n = 8$) and mean pupil dilation per task ($n = 8$) were specified as level 1 dependent variables (i.e., tasks within participants design).

4.5.2. *Hypothesis 5*

To investigate whether time to locate relevant structure predicts cognitive load, for both dependent variables the same approach was taken. First an intercept only model was specified to establish the variance in the dependent variable explained by individual differences (level 2 variance) and variance explained by task differences (level 1 variance). Subsequently time to locate relevant structure was entered as a level 1 predictor of the dependent variable and its significance examined. Subsequently, the decrease in level 1 variance in the second model compared to the intercept only model was examined to gain insight into the explained variance in cognitive load measures by time to locate relevant structure.

5. Results Study 2

5.1. *Assumptions*

On six occasions, the eye tracker did not measure participants correctly, resulting in missing values. Examination of the occurrence of erroneous measurement showed that the six cases were spread over the tasks and the participants. No outliers were identified.

5.2. *Descriptive statistics*

Descriptive statistics are reported in Table 4 concerning the eight participants.

Table 4
Descriptive statistics Study 2.

	M	*SD*	Valid *n*
Mean pupil dilation (all participants)	2.72	1.05	58
Participant 1	2.20	0.93	7
Participant 2	2.30	0.88	8
Participant 3	2.95	0.99	7
Participant 4	3.65	1.16	7
Participant 5	2.64	1.09	8
Participant 6	3.31	0.69	8
Participant 7	2.12	0.49	8
Participant 8	2.59	1.28	5
Time to locate relevant structure (all participants)	0.40	0.36	58
Participant 1	0.31	0.25	7
Participant 2	0.43	0.37	8
Participant 3	0.36	0.33	7
Participant 4	0.43	0.41	7
Participant 5	0.24	0.25	8
Participant 6	0.46	0.44	8
Participant 7	0.57	0.46	8
Participant 8	0.39	0.35	5

5.3. Hypothesis 5

To examine whether time to locate relevant structure predicted self-reported mental effort two multilevel models were specified with participants as the second-level grouping variable (see Table 5). The intercept only model (Model 1) had an intraclass correlation of <1%. This indicates that very little of the variance in self-reported mental effort could be attributed to differences between participants; most variance was attributed to task-differences. Subsequently, time to locate relevant structure was added to the model as a level 1 predictor (Model 2). The predictor approached significance ($p = 0.059$) and explained 5% of the level 1 variance.

Subsequently, two multilevel models were specified with participants as the second-level grouping variable (see Table 5) to assess whether time to locate relevant structure predicted pupil dilation. The intercept only model (Model 1) had an intraclass correlation of 0.20, indicating that 20% of the variance in pupil dilation was observed between participants, whereas 80% was attributed to task-differences. Subsequently, time to locate relevant structure was added to the model as a level 1 predictor (Model 2). The predictor showed to be significant ($p < 0.01$) and explained 17% of the first level variance, indicating that time to locate relevant structure contributed to variance in pupil dilation.

6. Discussion

Two studies were conducted to investigate how human-computer interaction and eye movements relate to cognitive load, and how cognitive load is affected in medical students engaged in volumetric image interpretation. In Study 1, three measures of human-computer interaction formed one latent variable volumetric image information that significantly predicted cognitive load. Cognitive load increased with increased human-computer interaction. Eye tracking results of students engaged in image interpretation in Study 2 showed that when a student takes longer to find a relevant structure, pupil dilation, as a proxy for cognitive load, significantly increased. However on self-reported mental effort, as another proxy for cognitive load, only a trend was observed.

The positive relationship of volumetric image information with self-reported mental effort in Study 1 suggests that human-computer interaction can capture factors that influence cognitive load, supporting its use as an indirect objective measure of cognitive load. Examining the specific human-computer interaction variables that volumetric image information is composed of (i.e., amount of slices displayed, number of angle changes, total time spent in task) demonstrated what aspects of the image interpretation task influence cognitive load. Higher scores on these variables point to more cognitive and perceptual processing of task information; it implies that more effort is put in searching (e.g., more scrolling, more angle changing, longer time to examine the images) and therefore there is greater exposure to visual information. That this relates to higher cognitive load is consistent with studies showing that novices, as opposed to experts, have to consciously consider separate features of medical images (Eva et al., 2002; Kundel, Nodine, Conant, & Weinstein, 2007; van Merriënboer & Sweller, 2010). These findings also provide further elaboration on research into simulated 3D environments and cognitive load (Schrader & Bastiaens, 2012; van der Land et al., 2013). van der Land et al. (2013) suggested that the increase in visual cues and interactivity in 3D environments is beneficial for individual understanding, but when additional factors come into play this may lead to higher cognitive load. The current study provides specific evidence for a positive relation between visual information, the interactivity it results from, and cognitive load in an individual setting, and shows how such measures could be used to determine what exactly causes the increased load. The positive relationship does not support the alternative expectation that more image information coming from the volumetric image could reduce students' cognitive load by providing more contextual information. Perhaps students' spatial skills mediate the relationship between cognitive load and volumetric image information. Stull et al. (2009) showed that high spatial ability individuals perform better than low spatial ability individuals in a 3D anatomy task. Interaction with 3D visualisations can however attenuate differences in spatial ability (Hegarty et al., 2007), highlighting the complexity of this issue. Including spatial skills as a variable in subsequent research could shed more light on the direction of this relationship and the mechanisms involved.

Overall test-performance was included as a covariate in the model to account for confounding variance in cognitive load caused by the skill of the participant. However, the main effect of overall test-performance on self-reported mental effort was insignificant, suggesting that how well students performed on the full test did not affect their cognitive load in Task 1. Consequently, the subsequently included interaction effect of overall test-performance with volumetric image information on self-reported mental effort was insignificant as well. An explanation for the insignificant main effect of overall test performance may be that Task 1 was not discriminative enough between high- and low-performers, owing to the sample being too homogeneous as it only consisted of second-year medical students.

Two human-computer interaction variables, time taken to locate first relevant slice and proportion of time spent on relevant slices, did not contribute to the latent variable volumetric image information. This was unexpected as they indicate whether crucial visual information is displayed. It is possible that these measures of human-computer interaction do not provide enough information on where specifically the students are looking on the slices, i.e. spending time on a relevant slice does not imply that the relevant structure is examined. This suggests that in order for these

Table 5
Results multi-level analysis self-reported mental effort and pupil dilation on time to locate relevant structure.

	Self-reported mental effort				Pupil dilation			
	Model 1		Model 2		Model 1		Model 2	
	B	*SE*	*B*	*SE*	*B*	*SE*	*B*	*SE*
Intercept	5.07^{***}	0.27	5.07^{***}	0.27	2.64^{***}	0.21	2.64^{***}	0.20
Time to locate relevant structure			1.41^{*}	0.73			0.99^{**}	0.31
Variance								
Participant-level (2)	<0.01		<0.01		0.22^{**}		0.25^{**}	
Task-level (1)	4.28		4.07		0.86		0.72	

$^{***}p < 0.001$. $^{**}p < 0.01$. $^{*}p < 0.10$.

variables to be meaningful for the model, more precise data is needed. More precise data was available in Study 2; using eye tracking it was investigated what students examine in a medical image. In this study, time to locate a relevant structure was a significant predictor of pupil dilation. Although conclusions should be drawn with care as only a trend was observed on self-reported mental effort, this seems in line with research in other contexts where relevant information, and specifically the time taken to locate it, was related to cognitive load (De Fockert et al., 2001; Lavie, 2005).

6.1. Implications for cognitive load research

The current research contributes to a growing body of research using indirect and direct objective measures to measure cognitive load such as EEG (Antonenko et al., 2010), skin response (Nourbakhsh et al., 2013), speech (Khawaja, Chen, & Marcus, 2014), eye movements (Chen & Epps, 2013) and human-computer interaction (Ruiz et al., 2011, 2007). The use of such measures is advocated to gain a direct insight into what factors contribute to cognitive load (Martin, 2014; Sweller et al., 2011). A next step would be to integrate the relevant human-computer interaction and eye movement variables established in the current research into one model, to identify unique variance in perceived cognitive load accounted for by each variable. An extension to the proposed model would be to include a more discriminating measure of performance. No evidence was found for a performance effect on cognitive load in this study, but the use of a relatively homogeneous sample here combined with theoretical as well as empirical work strongly supporting a performance effect (Eva et al., 2002; van Gog et al., 2009) suggests this may be a methodological issue. One variable of specific interest in a model including performance is time to locate relevant information. A relation between examination of relevant information with performance and expertise has been firmly established (van Gog et al., 2009), and inclusion of such a variable would allow exploration of any mediating and moderating relationships between time to locate relevant information, performance, and cognitive load.

Previous research into usability and human-computer interaction on cognitive load has also focused on the load caused by the interface of software (Hollender et al., 2010). I.e. a non-intuitive interface requires an individual to direct cognitive resources away from the task at hand and towards operating the interface. Such cognitive load is not beneficial to learning (i.e. extraneous cognitive load) and should therefore be avoided. A limitation of the current study is that no distinction has been made between different types of cognitive load, although it is conceivable that part of the cognitive load experienced is in fact extraneous due to interaction with the software. Disentangling different types of load has in the past proved elusive (Martin, 2014), but there is evidence that different objective measures tap into different types of load (Antonenko & Niederhauser, 2010; Zheng & Cook, 2012). In the current context, in particular eye tracking is promising as it indicates how much attention is given to the task and how much to the software interface, and therefore allows measures of each type to be extracted from the data. As such, distinguishing between different types of load in the current proposed model is a relevant avenue for future investigation.

6.2. Implications for medical education

This research provides initial insights into how the volumetric image interpretation task affects cognitive load. The increased cognitive load due to more human-computer interaction may yield a positive influence on the development of image interpretation skills, as the measured activity suggests engagement with the task and therefore task-relevant information is processed in working memory (Hollender et al., 2010; Sweller, 2004). Too much cognitive load however has proven to be detrimental to learning performance in other contexts (van Merriënboer & Sweller, 2010) and should be avoided. When designing the curriculum a careful consideration between volumetric images which are potentially more demanding but better resembling medical practice, and 2D images which are low in image information but can be useful in teaching basic skills, should be made. Additionally, attention for usability of the software used for volumetric image interpretation as well as appropriate training in use of the software have proven to be useful for managing cognitive load (Clarke, Ayres, & Sweller, 2005; Hollender et al., 2010). Finally, tailored guidance (i.e., scaffolding, see Grunwald & Corsbie-Massay, 2006) to support task completion and comprehension could be employed in volumetric image interpretation to assist in lowering cognitive load. The established human-computer interaction and eye movement measures of cognitive load could inform such interventions, but it is important to note that an effect of learning image interpretation skills was not investigated in this study. Previous research has shown the benefits of volumetric image interpretation on the development of image interpretation skills, but has also demonstrated reduced performance of students when compared to 2D image interpretation (Ravesloot, van der Gijp, et al., 2015; Ravesloot, van der Schaaf, et al., 2015). Investigation of the relation between cognitive load and development of image interpretation skills would be valuable to provide specific direction to the aforementioned interventions.

6.3. Conclusion

This explorative study has contributed to evidence on how human-computer interaction and eye movements are related to cognitive load, and initial insights have been obtained into how cognitive load is affected in volumetric image interpretation. It is argued that by combining the human-computer interaction and eye movement variables comprehensive indirect objective measurement of cognitive load can occur. Combining such a model with investigation of learning effects of volumetric image interpretation in medical education may lead to relevant recommendations for implementation in the curriculum.

Acknowledgements

The research was partly financially supported by SURF Foundation, Collaborative organisation for ICT in Dutch higher education and research. SURF had no involvement in the study design, analysis, interpretation of the data or drafting of the article. The authors are very grateful to Ignace Hooge for providing us with his experience and expertise on the processing of eye tracking data. The authors are also very grateful to Josephine Huige for her very well organised support during the data-collection process.

References

Andriole, K. P., Wolfe, J. M., Khorasani, R., Treves, S. T., Getty, D. J., Jacobson, F. L., … Seltzer, S. E. (2011). Optimizing analysis, visualization, and navigation of large image data sets: one 5000-section CT scan can ruin your whole day. *Radiology, 259*(2), 346–362. http://doi.org/10.1148/radiol.11091276.

Antonenko, P. D., & Niederhauser, D. S. (2010). The influence of leads on cognitive load and learning in a hypertext environment. *Computers in Human Behavior, 26*(2), 140–150. http://doi.org/10.1016/j.chb.2009.10.014.

Antonenko, P. D., Paas, F., Grabner, R., & van Gog, T. (2010). Using electroencephalography to measure cognitive load. *Educational Psychology Review, 22*(4), 425–438. http://doi.org/10.1007/s10648-010-9130-y.

Bailey, B. P., & Iqbal, S. T. (2008). Understanding changes in mental workload during execution of goal-directed tasks and its application for interruption management. *ACM Transactions on Computer-Human Interaction (TOCHI), 14*(4), 1–28. http://doi.org/10.1145/1314683.1314689.

Brünken, R., Plass, J. L., & Leutner, D. (2003). Direct measurement of cognitive load in multimedia learning. *Educational Psychologist, 38*(1), 53–61. http://doi.org/10.1207/S15326985EP3801_7.

Brünken, R., Steinbacher, S., Schnotz, W., Plass, J. L., & Leutner, D. (2002). Assessment of cognitive load in multimedia learning using dual-task methodology. *Experimental Psychology, 49*(2), 109–119. http://doi.org/10.1027//1618-3169.49.2.109.

Byrne, B. M. (2012). *Structural equation modeling with MPlus. Basic concepts, applications, and programming.* New York, NY: Routledge.

Chen, S., & Epps, J. (2013). Automatic classification of eye activity for cognitive load measurement with emotion interference. *Computer Methods and Programs in Biomedicine, 110*(2), 111–124. http://doi.org/10.1016/j.cmpb.2012.10.021.

Chi, M. T. H., Glaser, R., & Rees, E. (1982). Expertise in problem solving. In R. Sternberg (Ed.), *Advances in the psychology of human intelligence* (pp. 7–75). Hillsdale, NJ: Erlbaum.

Clarke, T., Ayres, P., & Sweller, J. (2005). The impact of sequencing and prior knowledge on learning mathematics through spreadsheet applications. *Educational Technology Research and Development, 53*(3), 15–24. http://doi.org/10.1007/BF02504794.

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience, 3*(3), 201–215. http://doi.org/10.1038/nrn755.

De Fockert, J. W., Rees, G., Frith, C. D., & Lavie, N. (2001). The role of working memory in visual selective attention. *Science, 291*(5509), 1803–1806. http://doi.org/10.1126/science.1056496.

DeLeeuw, K. E., & Mayer, R. E. (2008). A comparison of three measures of cognitive load: evidence for separable measures of intrinsic, extraneous, and germane load. *Journal of Educational Psychology, 100*(1), 223–234. http://doi.org/10.1037/0022-0663.100.1.223.

Donald, J. J., & Barnard, S. A. (2012). Common patterns in 558 diagnostic radiology errors. *Journal of Medical Imaging and Radiation Oncology, 56*(2), 173–178. http://doi.org/10.1111/j.1754-9485.2012.02348.x.

Ellis, S. M., Hu, X., Dempere-Marco, L., Yang, G. Z., Wells, A. U., & Hansell, D. M. (2006). Thin-section CT of the lungs: eye-tracking analysis of the visual approach to reading tiled and stacked display formats. *European Journal of Radiology, 59*(2), 257–264. http://doi.org/10.1016/j.ejrad.2006.05.006.

Eva, K. W., Norman, G. R., Neville, A. J., Wood, T. J., & Brooks, L. R. (2002). Expert/novice differences in memory: a reformulation. *Teaching and Learning in Medicine, 14*(4), 257–263. http://doi.org/10.1207/S15328015TLM1404_10.

Evers, A., Lucassen, W., Meijer, R., & Sijtsma, K. (2010). *COTAN beoordelingssysteem voor de kwaliteit van tests (COTAN ratingsystem for the quality of tests)*. Retrieved from http://dare.uva.nl/document/179627.

Fan, X., Thompson, B., & Wang, L. (1999). Effects of sample size, estimation methods, and model specification on structural equation modeling fit indexes. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 56–83. http://doi.org/10.1080/10705519909540119.

van Gerven, P. W. M., Paas, F. G. W. C., van Merriënboer, J. J. G., & Schmidt, H. G. (2004). Memory load and the cognitive pupillary response in aging. *Psychophysiology, 41*(2), 167–174. http://doi.org/10.1111/j.1469-8986.2003.00148.x.

van der Gijp, A., Ravesloot, C. J., van der Schaaf, M. F., van der Schaaf, I. C., Huige, J. C. B. M., Vincken, K. L., … van Schaik, J. P. J. (2015). Volumetric and two-dimensional image interpretation show different cognitive processes in learners. *Academic Radiology, 22*(5), 632–639. http://doi.org/10.1016/j.acra.2015.01.001.

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 in Human Behavior, 25*(2), 325–331. http://doi.org/10.1016/j.chb.2008.12.021.

Grunwald, T., & Corsbie-Massay, C. (2006). Guidelines for cognitively efficient multimedia learning tools: educational strategies, cognitive load, and interface design. *Academic Medicine, 81*(3), 213–223. http://doi.org/10.1097/00001888-200603000-00003.

Hegarty, M., Keehner, M., Cohen, C., Montello, D. R., & Lippa, Y. (2007). The role of spatial cognition in medicine: applications for selecting and training professionals. In G. Allan (Ed.), *Applied spatial cognition*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hollender, N., Hofmann, C., Deneke, M., & Schmitz, B. (2010). Integrating cognitive load theory and concepts of human-computer interaction. *Computers in Human Behavior, 26*(6), 1278–1288. http://doi.org/10.1016/j.chb.2010.05.031.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1–55. http://doi.org/10.1080/10705519909540118.

Khawaja, M. A., Chen, F., & Marcus, N. (2014). Measuring cognitive load using linguistic features: implications for usability evaluation and adaptive interaction design. *International Journal of Human-Computer Interaction, 30*(5), 343–368. http://doi.org/10.1080/10447318.2013.860579.

Kirschner, F., Paas, F. G. W. C., & Kirschner, P. A. (2009). Individual and group-based learning from complex cognitive tasks: effects on retention and transfer efficiency. *Computers in Human Behavior, 25*(2), 306–314. http://doi.org/10.1016/j.chb.2008.12.008.

Koh, K. H., & Zumbo, B. D. (2008). Multi-group confirmatory factor analysis for testing measurement invariance in mixed item format data. *Journal of Modern Applied Statistical Methods, 7*(2), 471–477.

Kok, E. M., Jarodzka, H., de Bruin, A. B. H., BinAmir, H. A. N., Robben, S. G. F., & van Merrienboer, J. J. G. (2015). Systematic viewing in radiology: seeing more, missing less? *Advances in Health Sciences Education, 21*(1), 189–205. http://doi.org/10.1007/s10459-015-9624-y.

Krupinski, E. A. (2010). Current perspectives in medical image perception. *Attention, Perception, and Psychophysics, 72*(5), 1205–1217. http://doi.org/10.3758/APP.72.5.1205.

Krupinski, E. A. (2011). The role of perception in imaging: past and future. *Seminars in Nuclear Medicine, 41*(6), 392–400. http://doi.org/10.1053/j.semnuclmed.2011.05.002.

Krupinski, E. A., Berbaum, K. S., Caldwell, R. T., Schartz, K. M., Madsen, M. T., & Kramer, D. J. (2012). Do long radiology workdays affect nodule detection in dynamic CT interpretation? *Journal of the American College of Radiology, 9*(3), 191–198. http://doi.org/10.1016/j.jacr.2011.11.013.

Kundel, H. L., Nodine, C. F., Conant, E. F., & Weinstein, S. P. (2007). Holistic component of image perception in mammogram interpretation: gaze-tracking study. *Radiology, 242*(2), 396–402. http://doi.org/10.1148/radiol.2422051997.

van der Land, S., Schouten, A. P., Feldberg, F., van den Hooff, B., & Huysman, M. (2013). Lost in space? Cognitive fit and cognitive load in 3D virtual environments. *Computers in Human Behavior, 29*(3), 1054–1064. http://doi.org/10.1016/j.chb.2012.09.006.

Lavie, N. (2005). Distracted and confused?: Selective attention under load. *Trends in Cognitive Sciences, 9*(2), 75–82. http://doi.org/10.1016/j.tics.2004.12.004.

Lesgold, A. M., Rubinson, H., Feltovich, P., Glaser, R., Klopfer, D., & Wang, Y. (1988). Expertise in a complex skill: diagnosing x-ray pictures. In M. T. H. Chi, R. Glaser, & J. Farr (Eds.), *The nature of expertise* (pp. 311–342). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Martin, S. (2014). Measuring cognitive load and cognition: metrics for technology-enhanced learning. *Educational Research and Evaluation, 20*(7–8), 592–621. http://doi.org/10.1080/13803611.2014.997140.

Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist, 38*(1), 43–52. http://doi.org/10.1207/S15326985EP3801_6.

van Merriënboer, J. J. G., & Sweller, J. (2005). Cognitive load theory and complex learning: recent developments and future directions. *Educational Psychology Review, 17*(2), 147–177. http://doi.org/10.1007/s10648-005-3951-0.

van Merriënboer, J. J. G., & Sweller, J. (2010). Cognitive load theory in health professional education: design principles and strategies. *Medical Education, 44*(1), 85–93. http://doi.org/10.1111/j.1365-2923.2009.03498.x.

Norman, G. R., Coblentz, C. L., Brooks, L. R., & Babcook, C. J. (1992). Expertise in visual diagnosis: a review of the literature. *Academic Medicine, 67*(10), 78–83. http://doi.org/10.1097/00001888-199210000-00045.

Nourbakhsh, N., Wang, Y., & Chen, F. (2013). GSR and blink features for cognitive load classification. In P. Kotzé, G. Marsden, G. Lindgaard, J. Wesson, & M. Winckler (Eds.), *Human-computer interaction – INTERACT 2013* (pp. 159–166). Berlin, Heidelberg, Germany: Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-642-40483-2_11.

Paas, F. G. W. C. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: a cognitive load approach. *Journal of Educational Psychology, 84*(4), 429–434. http://doi.org/10.1037//0022-0663.84.4.429.

Paas, F. G. W. C., Tuovinen, J. E., Tabbers, H., & van Gerven, P. W. M. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist, 38*(1), 63–72. http://doi.org/10.1207/S15326985EP3801_8.

Paas, F. G. W. C., van Merriënboer, J. J. G., & Adam, J. J. (1994). Measurement of cognitive load in instructional research. *Perceptual & Motor Skills, 79*(1), 419–430. http://doi.org/10.2466/pms.1994.79.1.419.

Palinko, O., Kun, A. L., Shyrokov, A., & Heeman, P. (2010). Estimating cognitive load using remote eye tracking in a driving simulator. In *Symposium on Eye-Tracking Research & Applications* (pp. 141–144). http://dx.doi.org/10.1145/1743666.1743701.

Pinto, A., Acampora, C., Pinto, F., Kourdioukova, E., Romano, L., & Verstraete, K. (2011). Learning from diagnostic errors: a good way to improve education in radiology. *European Journal of Radiology, 78*(3), 372–376. http://doi.org/10.1016/j.ejrad.2010.12.028.

Ravesloot, C. J., van der Gijp, A., van der Schaaf, M. F., Huige, J. C. B. M., Vincken, K. L., Mol, C. P., … van Schaik, J. P. J. (2015). Support for external validity of radiological anatomy tests using volumetric images. *Academic Radiology, 22*(5), 640–645. http://doi.org/10.1016/j.acra.2014.12.013.

Ravesloot, C. J., van der Schaaf, M. F., van Schaik, J. P. J., ten Cate, O. T. J., van der Gijp, A., Mol, C. P., et al. (2015). Volumetric CT-images improve testing of radiological image interpretation skills. *European Journal of Radiology, 84*(5), 856–861. http://dx.doi.org/10.1016/j.ejrad.2014.12.015.

Reiner, B. I., Siegel, M. D., & Siddiqui, K. (2003). Evolution of the digital revolution: a radiologist perspective. *Journal of Digital Imaging, 16*(4), 324–330. http://doi.org/10.1007/s10278-003-1743-y.

Rengier, F., Häfner, M. F., Unterhinninghofen, R., Nawrotzki, R., Kirsch, J., Kauczor, H.-U., et al. (2013). Integration of interactive three-dimensional image post-processing software into undergraduate radiology education effectively improves diagnostic skills and visual-spatial ability. *European Journal of Radiology, 82*(8), 1366–1371. http://doi.org/10.1016/j.ejrad.2013.01.010.

Ruiz, N., Taib, R., & Chen, F. (2011). Freeform pen-input As evidence of cognitive load and expertise. In *Proceedings of the 13th International Conference on Multimodal Interfaces* (pp. 185–188). New York, NY: ACM. http://doi.org/10.1145/2070481.2070511.

Ruiz, N., Taib, R., Shi, Y.(D.), Choi, E., & Chen, F. (2007). Using pen input features as indices of cognitive load. In *Proceedings of the 9th International Conference on Multimodal Interfaces* (pp. 315–318). New York, NY: ACM. http://doi.org/10.1145/1322192.1322246.

Satorra, A. (2000). Scaled and adjusted restricted tests in multi-sample analysis of moment structures. In R. D. H. Heijmans, D. S. G. Pollock, & A. Satorra (Eds.), *Innovations in multivariate statistical analysis. A festschrift for Heinz Neudecke* (pp. 223–247). London, UK: Kluwer Academic Publishers. http://doi.org/10.1007/978-1-4615-4603-0_17.

Schrader, C., & Bastiaens, T. J. (2012). The influence of virtual presence: effects on experienced cognitive load and learning outcomes in educational computer games. *Computers in Human Behavior, 28*(2), 648–658. http://doi.org/10.1016/j.chb.2011.11.011.

Stull, A. T., Hegarty, M., & Mayer, R. E. (2009). Getting a handle on learning anatomy with interactive three-dimensional graphics. *Journal of Educational Psychology, 101*(4), 801–816. http://doi.org/10.1037/a0016849.

Sweller, J. (2004). Instructional design consequences of an analogy between evolution by natural selection and human cognitive architecture. *Instructional Science, 32*(1–2), 9–31. http://dx.doi.org/10.1023/B:TRUC.0000021808.72598.4d.

Sweller, J., Ayres, P., & Kalyuga, S. (2011). Measuring cognitive load. *Cognitive load theory: Explorations in the learning sciences, instructional systems and performance technologies* (pp. 71–85). New York, NY: Springer. http://doi.org/10.1007/978-1-4419-8126-4_6.

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. http://doi.org/10.1023/A:1022193728205.

Taylor, P. M. (2007). A review of research into the development of radiologic expertise: implications for computer-based training. *Academic Radiology, 14*(10), 1252–1263. http://doi.org/10.1016/j.acra.2007.06.016.

Tobii technology A. B.. (2010). *Tobii eye tracking: An introduction to eye tracking and Tobii eye trackers*. Retrieved from http://www.tobii.com/Global/Analysis/Training/WhitePapers/Tobii_EyeTracking_Introduction_WhitePaper.pdf?epslanguage=en.

Vincken, K. L., & Ravesloot, C. J. (2010). *VQuest*. Utrecht, The Netherlands: ISI UMCU.

Zheng, R., & Cook, A. (2012). Solving complex problems: a convergent approach to cognitive load measurement. *British Journal of Educational Technology, 43*(2), 233–246. http://doi.org/10.1111/j.1467-8535.2010.01169.x.