Computers in Human Behavior 99 (2019) 268–277

Contents lists available at ScienceDirect

Computers in Human Behavior

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

Full length article

How prior knowledge affects problem-solving performance in a medical simulation game: Using game-logs and eye-tracking

Joy Yeonjoo Lee[a,*], Jeroen Donkers[a], Halszka Jarodzka[b], Jeroen J.G. van Merriënboer[a]

[a] School of Health Professions Education, Maastricht University, P.O. Box 616, 6200, MD, Maastricht, the Netherlands
[b] Welten Institute, Open University of The Netherlands, PO Box 2960, 6401, DL, Heerlen, the Netherlands

ARTICLE INFO

Keywords:
Assessment
Prior knowledge
Simulation game
Serious game
Cognitive load
Eye-tracking

ABSTRACT

Computer-based simulation games provide an environment to train complex problem-solving skills. Yet, it is largely unknown how the in-game performance of learners varies with different levels of prior knowledge. Based on theories of complex-skill acquisition (e.g., 4C/ID), we derive four performance aspects that prior knowledge may affect: (1) systematicity in approach, (2) accuracy in visual attention and motor reactions, (3) speed in performance, and (4) cognitive load. This study aims to empirically test whether prior knowledge affects these four aspects of performance in a medical simulation game for resuscitation skills training. Participants were 24 medical professionals (experts, with high prior knowledge) and 22 medical students (novices, with low prior knowledge). After pre-training, they all played one scenario, during which game-logs and eye-movements were collected. A cognitive-load questionnaire ensued. During game play, experts demonstrated a more systematic approach, higher accuracy in visual selection and motor reaction, and a higher performance speed than novices. Their reported levels of cognitive load were lower. These results indicate that prior knowledge has a substantial impact on performance in simulation games, opening up the possibility of using our measures for performance assessment.

1. Introduction

Computer-based simulation games (CBSG) are effective learning environments for complex skills. As simulations, they approximately replicate the complexity of real-life situations (Koivisto, Niemi, Multisilta, & Eriksson, 2017). As computer games, they provide a package of problems that are causally connected, based on learners' interaction with the game (Kiili, 2005). In this simulated problem-solving environment, learners can train specific professional skills in areas such as aviation, business management, and medicine (Dankbaar et al., 2016; De Freitas, 2006; Hernández-Lara, Perera-Lluna, & Serradell-López, 2019). However, CBSGs face a challenge in that the performance of a learner in the game is difficult to assess via traditional measurements such as achievement tests (Kang, Liu, & Qu, 2017). This challenge is mainly due to the open-ended nature of CBSGs (Squire, 2008), which allows for a large number of different behaviors. Therefore, recent research has focused on tracking users' in-game behaviors by looking at game data such as *serious game analytics* (Kang et al., 2017; Loh, Sheng, & Ifenthaler, 2015; Wallner & Kriglstein, 2013). These studies identified several limitations: Data analysis without involving educational theoretical principles often fails to fully account for students' performance (Kang et al., 2017), game-logs without translation into high-level meaningful actions can yield confounding information (Zhou, Xu, Nesbit, & Winne, 2010), some important factors such as timing cannot be explained by analyzing sequences of events only (Clark, Martinez-Garza, Biswas, Luecht, & Sengupta, 2012), and empirical studies about how game data can be informative for performance assessment are scarce (Hou, 2015; Kang et al., 2017).

We believe that theories of complex-skill acquisition might help to develop performance assessments in open-ended game environments. We view the playing of a CBSG as a problem-solving process in which domain-specific prior knowledge (DSPK) has an essential role. DSPK comprises knowledge structures in long-term memory, also known as cognitive schemas (Bartlett, 1995). Without these schemas, learners depend on domain-general problem-solving strategies which are inefficient and time-consuming and, most importantly, hamper the schema construction processes (Van Merriënboer, 2013). This means that playing a CBSG without sufficient DSPK might lead to suboptimal learning. The goal of this study is to empirically examine the effect of DSPK on game performance by comparing learners with two distinct

* Corresponding author.,
E-mail addresses: joy.lee@maastrichtuniversity.nl (J.Y. Lee), jeroen.donkers@maastrichtuniversity.nl (J. Donkers), Halszka.Jarodzka@ou.nl (H. Jarodzka), j.vanmerrienboer@maastrichtuniversity.nl (J.J.G. van Merriënboer).

https://doi.org/10.1016/j.chb.2019.05.035
Received 13 February 2019; Received in revised form 30 April 2019; Accepted 29 May 2019
Available online 30 May 2019

levels of DSPK: learners with high DSPK (i.e., experts) and learners with low DSPK (i.e., novices).

Expert-novice differences in the use of problem-solving strategies have been investigated in multiple studies, suggesting various indicators of these differences (Donovan & Litchfield, 2013; Manning, Ethell, Donovan, & Crawford, 2006; McLaughlin, Bond, Hughes, McConnell, & McFadden, 2017). However, these indicators are highly conditional because problem-solving strategies greatly vary depending on domains and task environments (Ericsson, Hoffman, Kozbelt, & Williams, 2018; Liversedge, Gilchrist, & Everling, 2011, Chapter 30). To make indicators informative to education, they should be developed within an integral theoretical framework in education, specialized to a given task environment via careful task analysis, and validated by empirical studies. Regarding that the task environments of CBSGs are exceedingly dynamic and the tasks require interactions of performers with the environment, this study demonstrates how to develop specific indicators for a CBSG based on complex-skill acquisition theories.

In this introduction, we will first theoretically compare how experts and novices generate problem solutions, suggesting aspects of problem-solving performance that are directly affected by the level of DSPK. We will then discuss how to define indicators of these aspects by decomposing the skill structure hierarchically. Finally, we present the hypotheses of this study.

1.1. Problem solution generation by experts and novices

Fig. 1 provides a process model that shows how experts and novices generate problem solutions differently, adopting concepts from the four-component instructional design (4C/ID) model (Van Merriënboer & Kirschner, 2018). The process involves two types of knowledge in long-term memory: *domain models* (i.e., schemas of how a domain is organized) at *declarative level*, and *cognitive strategies* (i.e., schemas of how to approach problems in the domain) at *procedural level*. Assume a continuum with novices with low DSPK at one extreme and experts with high DSPK at the other extreme. For novices, since their domain models are not yet structured, *weak methods* (i.e., slow and inefficient general problem-solving strategies such as *general search* or *working backward*) (Newell & Simon, 1972) are the only cognitive strategies they can use when solving a problem. This leads to *inefficient approaches* to the problem, and also to *procedures with incorrect cognitive rules* at the level of task performance (i.e., *solution generated*). For experts, on the other hand, well-structured domain models are interpreted and transformed into two types of stronger cognitive strategies: *knowledge-based methods* (i.e., heuristic strategies) and *strong methods* (i.e., algorithmic strategies) (Van Merriënboer, 2013). Knowledge-based methods guide students to reason within the domain and systematically approach *non-routine* aspects of the problem (i.e., *systematic approach*). When a certain aspect of the given task is consistently repeated (i.e., *routine* aspects of tasks), cognitive if-then rules may be formed as strong methods. These rules provide algorithmic solutions to routine aspects of the task by matching conditional information in working memory (i.e., *if* part) with a coordinated reaction (i.e., *then* part), resulting in *procedures with correct cognitive rules* at the task performance level. As a function of extensive practice, the cognitive rules can be strengthened and eventually become fully automatized, leading to higher speed in performance (Palmeri, 1999).

Additionally, the schemas embodied in long-term memory cause one more distinction in task performance between experts and novices: reduced cognitive load resulting from optimized use of working memory. Problem-solving with weak methods imposes a heavy demand on cognitive resources in working memory (Sweller, Clark, & Kirschner, 2010), introducing high cognitive load or even cognitive overload (Sweller, 1988). However, with the availability of knowledge-based methods, cognitive schemas relevant for problem-solving are stored in long-term memory and retrieved into working memory as one element. Moreover, with fully automatized strong methods, cognitive schemas are activated directly without placing any demand on working memory resources, which further frees up working memory (Sweller, van Merriënboer, & Paas, 2019).

Consequently, we derive four constructs that represent aspects of task performance that are affected by DSPK: (1) systematicity in task approach (i.e., representation of acquired strategies), (2) accuracy in applying cognitive if-then rules (i.e., representation of formed cognitive rules), (3) speed in performance (i.e., representation of the strength of those rules), and (4) reduced level of cognitive load (i.e., representation of optimized process).

For example, in emergency medicine, an expert with knowledge and strategies in the domain would approach an emergency case

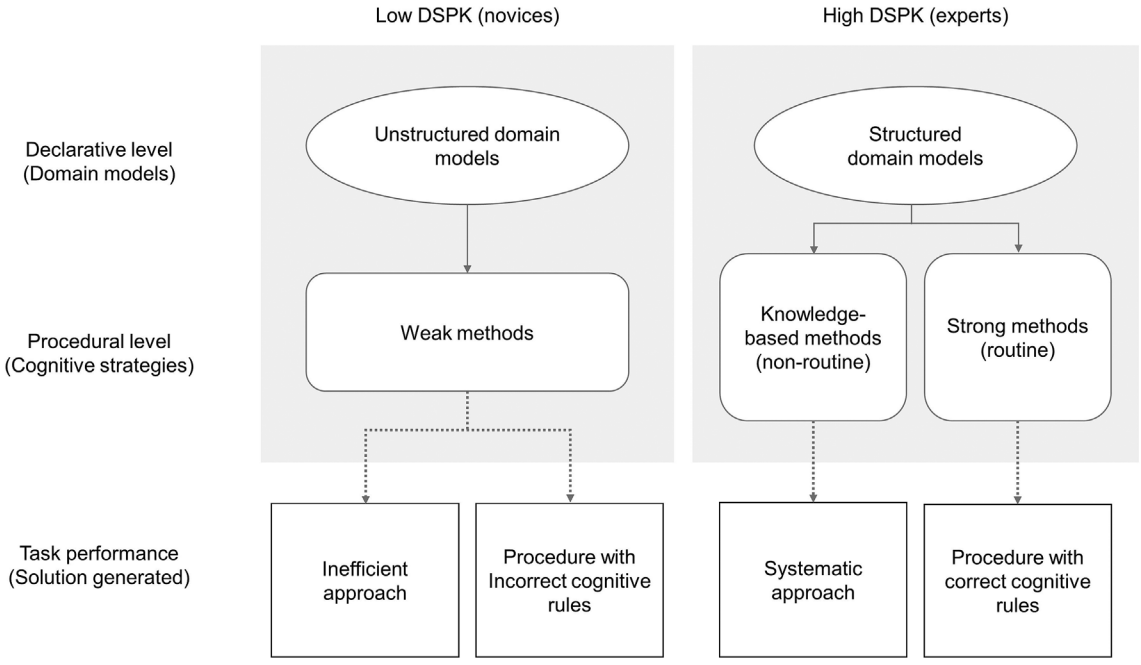

Fig. 1. The model of problem solution generation by experts and novices.

systematically by reasoning in terms of priorities of interventions (systematicity in task approach). As for algorithmic rule-based aspects of the case (e.g., if the patient's oxygen level decreases, then apply oxygenation), the expert would perform the rule without errors (accuracy in applying cognitive rules). Since the expert has extensively practiced this rule, speed should be high (speed in performance). In addition, the expert would experience low cognitive load because knowledge from long-term memory can be applied directly (reduced level of cognitive load).

1.2. Skill decomposition

While the four aspects of task performance are applicable to all task environments, *indicators* to assess these aspects will be specific to a particular task environment. Researchers have strongly recommended that, to assess a certain task performance, constituent skills and their relationships should be identified in a process of skill decomposition (Gagne, 1968; Van Merriënboer & Kirschner, 2018). We deem that skill decomposition allows identification of the indicators of the four constructs mentioned above to be precise and theoretically sound.

The domain of this study is a resuscitation procedure, called the ABCDE method. The five letters ABCDE represent the five phases (i.e., Airway, Breathing, Circulation, Disabilities, Exposure) that a task performer goes through sequentially to stabilize an acutely ill patient. The sequence should be rigidly followed, based on the principle “treat first what kills first”. AbcdeSIM (Erasmus University Medical Center & VirtualMedSchool, 2012), a CBSG for training the ABCDE method, is employed as the task environment. We decompose the task and develop a skill hierarchy by using Lee and Anderson's task analysis method (Lee & Anderson, 2001) (Fig. 2). In the hierarchy, the *task-goal* (i.e., stabilization of patient) is gradually divided into three levels: unit-task level, functional level, and physical level. To achieve the task-goal, *unit-tasks* (i.e., the five phases in the ABCDE method) are arranged accordingly. Each unit-task comprises multiple sub-tasks at the *functional level* (i.e., diagnosis and intervention). Every functional task is linked to individual activities at the *physical level* (e.g., look at “VFM”, click “Talk to nurse”). What one can empirically measure is this physical level only, while other levels represent cognitive performance.

This hierarchy guides us in the development of the indicators of the four constructs, by identifying different task levels. For the first construct (systematicity), a systematic approach in the ABCDE principle can be defined as a high level of adherence to the order of the five phases at the unit-task level. The challenge of measuring this construct is that the systematic approach is not directly observable at the physical level. To see this, note that the knowledge-based methods deal with non-routine aspects of a task, using the *same* knowledge *differently* based on rules-of-thumb (Van Merriënboer, 2013). An action that is associated with a certain ABCDE phase can also be taken during other phases strategically. Thus, an irregular ABCDE sequence observed at the physical level does not necessarily represent irregular performance at other levels. Consequently, the indicator of this construct should concern the hidden cognitive performance at the unit-task level, rather than analyzing the physical level only.

For the second construct (accuracy in applying cognitive rules), we recall that cognitive rules consist of *if* and *then* parts. In CBSGs in general, the *if* part emerges as information gathering via visual selection (i.e., looking at a particular part of the screen), while the *then* part corresponds to motor reaction to the task environment (e.g., mouse clicks or keyboard input). The accuracy in visual selection and motor reaction might reside at the functional level in the skill hierarchy. This construct can be directly detected by observing the physical level, because the strong methods deal with routine aspects of a task, referring to the *same* use of *same* knowledge (Van Merriënboer, 2013).

The third construct, the strength of cognitive rules, is situated in the connection between the sub-tasks at the functional level. If this connection is strong and stable, certain motor reactions at the end of a series of cognitive rules will be performed fast. The indicator of the strength should be the speed of this motor reaction, observed again at the physical level.

Lastly, the construct of reduced cognitive load originates from well-structured cognitive schemas. The entire structure of the skill hierarchy shows how the relevant cognitive schemas in long-term memory are developed, resulting in optimization of use of working memory. The degree of optimization can be indicated by the level of cognitive load.

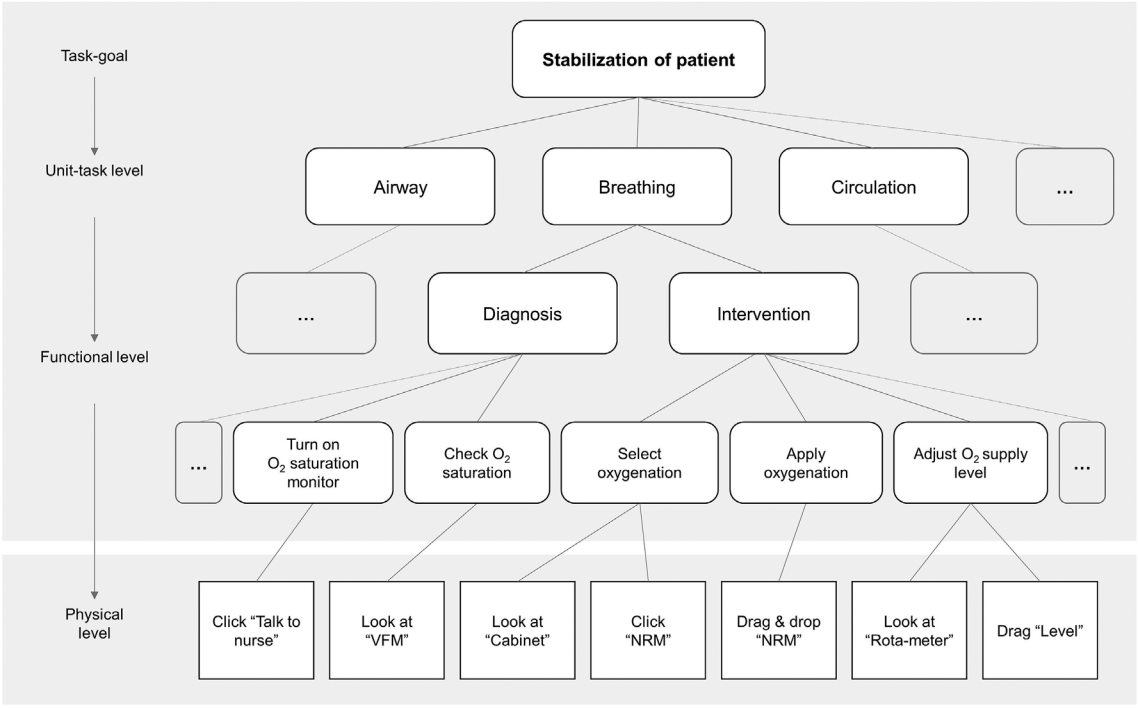

Fig. 2. Skill hierarchy of AbcdeSIM. The task-goal is divided into three levels: unit-task level, functional level, and physical level.

Table 1
Constructs, hypotheses, and indicators.

Construct	Hypothesis	Indicators
Systematicity in approach (H1)	Experts adhere to the five ABCDE phases more rigidly than novices	• Hidden Markov model score that measures to what extent the order of the five phases was kept
Accuracy in visual selection (H2a)	Experts allocate more visual attention to critical diagnosis areas (CDA) than novices	• Proportion of dwell time in CDA • Proportion of fixation count in CDA • Proportion of fixation duration in CDA
Accuracy in motor reaction (H2b)	Experts complete functional tasks more accurately than novices	• Intervention completion score
Speed in performance (H3)	Experts complete functional tasks faster	• Time on intervention tasks
Cognitive load (H4)	Experts' richer and more automated schemas require less working memory capacity	• Cognitive load questionnaire • Average fixation duration • Fixation frequency • Transition rate

1.3. Hypotheses

Four hypotheses will be tested in the current study:

H1 (systematicity in approach). Participants with high DSPK (i.e., experts) show higher systematicity in approach than participants with low DSPK (i.e., novices) during performance in the AbcedSIM, demonstrating higher level of adherence to the ABCDE phases.

H2 (accuracy in applying cognitive rules). Experts show higher accuracy in visual selection by allocating more visual attention to critical diagnosis areas (H2a) and in motor reactions by completing more interventions (H2b) than novices.

H3 (speed in performance). Experts show higher speed in performance by completing interventions faster than novices.

H4 (reduced cognitive load). Experts experience lower cognitive load than novices.

Table 1 provides an overview of the constructs, hypotheses, and indicators. Details of indicators are described in the following method section.

2. Method

2.1. Participants and design

Participants (N = 46) were recruited on a voluntary basis from a medical center in the Netherlands. For the expert group, 24 residents in their second to fifth (final) year of residency training with an average of 3.1 years of experience in emergency departments (SD = 1.6) were recruited (Md = 29 years with a range from 26 to 44; 17 females). For the novice group, 22 medical students in their second to sixth academic year who had been taught the basics of emergency medicine but had received no training were recruited (Md = 23 years with a range from 20 to 26; 12 females). A causal-comparative design is adopted with the level of expertise as the single factor.

2.2. Material and apparatus

2.2.1. AbcdeSIM game set-up

AbcdeSIM is a medical simulation game to train the ABCDE method for resuscitation. The game starts with a storyline where users meet a virtual patient in an emergency room. The users are provided with tools for diagnosis (e.g., stethoscope, penlight) and intervention (e.g., infusion fluids, medication). Human physiology (e.g., respiration, circulation) of the patient is implemented in the game, giving feedback on user's interventions. A regular adult patient scenario, hemorrhagic shock due to gastrointestinal bleeding (GIB), was used. GIB is a scenario where learners should follow the basic ABCDE method, with most emphasis on the circulation phase. During the experiment, the game was run on a personal computer (Intel Core i7 2.67 GHz CPU, 1.98 GB RAM) and presented on a Dell 22 "LCD screen with a resolution of 1650 × 1080 pixels. Participants used a headset for sound effects and interaction with the simulation was done via the mouse.

2.2.2. Eye-tracking and game-log recording

The game log data, containing user-input (e.g., tools that participant used, actions taken), changes in the game (e.g., patient's physiological changes), and time stamps, were saved in JSON file format (www.json.org). Participants' eye movements were measured by an SMI RED remote eye-tracker (SensoMotoric Instruments GmbH, Teltow, Germany) with a sampling rate of 250 Hz. The SMI Experiment Center 3.5 software (version 3.2.11, www.smivision.com) was used to implement calibration, validation, stimulus presentation, and screen recording. Eye movement data was gathered via SMI iView X software (version 2.7.13).

2.2.3. Cognitive load questionnaire

The NASA Task Load Index (NASA-TLX) (Hart & Staveland, 1988) was used as a validated self-report questionnaire of cognitive load. It is a mental workload assessment tool for human-machine interaction domains such as aviation and aeronautics (Shamo, Dror, & Degani, 1999), healthcare (Weinger et al., 2000), and socio-technical fields (Grigg, Garrett, & Benson, 2012; Warm, Matthews, & Finomore Jr, 2017). The NASA-TLX provides an overall workload score with six subscales: mental demand, physical demand, temporal demand, performance, effort, and frustration. Certain wordings of the questionnaire were adapted to fit the game environment.

2.3. Procedure

Individual sessions were carried out in an eye-tracking laboratory at Maastricht University. First, participants were asked to sign an informed consent form and fill out a questionnaire about demographics and experience in emergency medicine. Then, a pre-training was provided to ensure that the level of game-specific knowledge (i.e., how to operate the game) was comparable between the expert and novice groups. After pre-training, additional time for participants to play around with a test scenario was given, to allow them to familiarize themselves with the game. When participants expressed their readiness, the GIB scenario was presented. The eye-tracking system was calibrated with a 9-point procedure, and validation followed directly. Participants had to stabilize the virtual patient in a maximum of 15 min, shown with a timer visible on screen during the entire session. As time pressure is an intrinsic component of cognitive load in medical emergencies, we controlled for the time pressure by measuring temporal demand that is one of the six subscales of NASA-TLX. After the scenario, participants filled out the NASA-TLX. The average time to complete a session was about 50 min.

2.4. Data analysis

For testing H1, H2b, and H3, the data from game-logs was used, while eye-tracking data was employed for testing H2a and H4. Parsing of the game-logs was performed using Python. Eye-tracking data of three experts and two novices were excluded due to low tracking ratio

below 85%. The average tracking ratio after the exclusion was 94.9%. Outliers for each measure were identified by Tukey boxplots and excluded. Statistical analysis for each construct was performed in R version 3.5.1 (R Development Core Team, 2010).

2.4.1. Systematicity in approach (H1)

We consider Hidden Markov Models (HMM) a suitable method to develop a score for measuring systematicity in approach, since they can be used to model *hidden state transitions* (i.e., phase arrangement at the unit-task level) based on a sequence of *emission states* (i.e., arrangement of motor reactions observed at the physical level) (Baum, Petrie, Soules, & Weiss, 1970). The probability structure resulting from fitting the HMM to participant data contains information about the level of the adherence to the ABCDE sequence in hidden states. We used this probability structure to compute our score for systematicity in approach.

To do this, first, we classified the functional tasks of the GIB scenario into each of the ABCDE phases. Then, user-input data relevant to these functional tasks was extracted from the raw data in the game log file. The extracted data comprises the emission state sequences of ABCDE for each participant. A HMM is fitted to the sequences, resulting in a probability structure with two matrices: a state transition probability matrix and an emission probability matrix. From these matrices, we calculated the HMM score by averaging the sum of the diagonal and upper co-diagonal in the state transition matrix and the diagonal sum of emission probability matrix (see Appendix for a complete explanation of the HMM score computation).

2.4.2. Accuracy in visual selection (H2a) and motor reaction (H2b)

Research in visual science reports that, compared to novices, experts allocate more attention to task-relevant than task-redundant areas (Gibson, 2014; Haider & Frensch, 1999; Reingold & Sheridan, 2011, p. 528). However, in a real-life simulation such as the AbcdeSIM, areas with information cannot simply be dichotomized as relevant versus redundant. Information and game functions are compactly organized within the limited area of the screen, and the level of relevance gradually differs. Thus, we categorized the reason the screen into four groups in consultation with a medical professional: critical diagnosis area with critically relevant information for diagnosis (CDA), non-critical diagnosis area with information relevant for diagnosis to some extent but not critical (NDA), intervention area with functions for intervention (IVA), and neutral area with additional functions such as connecting different information (NA) (Fig. 3). We hypothesized that experts allocate more attentional resources to CDA than novices, thus formulating H2a.

All area groups mentioned above comprised areas of interest (AOIs) forming the basis of the eye-tracking data analysis (Holmqvist et al., 2011, Chapter 6). Since the appearance and layout of these areas dynamically change according to users' input and activated game function, we adapted the AOIs accordingly. The raw eye-tracking data was analyzed by SMI BeGaze 3.6 software. Fixations were identified when the gaze velocity was less than 40 visual degrees per second, with a minimum duration of 50 ms.

Three eye-movement measures are employed: dwell time (gaze visiting time for an AOI from entry to exit), fixation count (number of fixations on an AOI), and fixation duration (time duration when the eye is relatively still at a position). Each measure is expected to capture different aspects of attentional resources. Dwell time indicates the time that a participant spent fixating on an AOI, where constituent metrics are not decomposed (Orquin & Holmqvist, 2018). Fixation count indicates frequency of reference to the stimulus (Orquin & Loose, 2013), while longer duration of fixations can mean a deeper cognitive processing (Holmqvist et al., 2011, Chapter 11). To make the measures comparable across participants, relative values for each AOI group were calculated: the dwell time for each AOI group was divided by total play time, while the fixation count was divided by total number of fixations during the entire scenario. The mean fixation duration for each AOI group was calculated.

Visual selection and its associated motor reaction cannot be matched one-to-one, due to the dynamic characteristic of CBSGs. Thus, the accuracy in motor reaction was operationalized independently from the visual selection. We hypothesized experts complete more intervention tasks than novices, formulating H2b. The intervention completion score was developed as follows. In consultation with a medical professional, we selected five intervention tasks that are theoretically essential in the GIB scenario: *oxygen mask application, fluid administration, blood administration, blood order,* and *calling gastroenterologists*. We then calculated the proportion of the intervention completed. This was done by extracting the corresponding data from the game log files.

2.4.3. Speed in performance (H3)

The relative time to complete the five intervention tasks from H2b was used as the speed measure. Clicking the game start button was taken as the start point, with clicking the button for applying one of the five interventions as the end point for that invention. We assume that the speed of intervention includes the speed of diagnosis as they are closely connected and performed simultaneously in emergency medicine. To make the time on each task comparable, z-scores were used. First, we checked whether the time-on-task per task was normally distributed. We then transformed each time-on-task into a z-score per task for each participant.

2.4.4. Cognitive load (H4)

In addition to using the cognitive load questionnaire (i.e., NASA-TLX) as a subjective rating, we used eye-tracking measurements as an objective indicator of cognitive load. Several studies have shown that high cognitive load is related with long fixation durations (Korbach, Brünken, & Park, 2016; Park, Knörzer, Plass, & Brünken, 2015) and high fixation frequency (Van Orden, Limbert, Makeig, & Jung, 2001; Van Orden, Nugent, La Fleur, & Moncho, 1998; Zelinsky, Rao, Hayhoe, & Ballard, 1997). We also included transition rate (i.e., the movement from one AOI to another per second) that has been used in several studies of working memory capacity (Holmqvist et al., 2011, Chapter 12). As cognitive load represents the level of optimization of working memory, we assume that a robust transition rate might be interpreted as an active cognitive process with an optimal use of working memory

Fig. 3. Areas of interest (AOI) definition: critical diagnosis area (CDA), non-critical diagnosis area (NDA), intervention area (IVA), and neutral area (NA).

Table 2
Outcomes for systematicity in approach, accuracy in motor reaction, speed in performance, and cognitive load.

Construct	Measure	Expert		Novice		*t* (*U*)	*df*	*p*	*d* (r)	95%CI
		M	*SD*	*M*	*SD*					
Systematicity in approach (H1)	HMM score	0.79	0.06	0.71	0.08	−3.49	32	.001[b]	−1.16	-.13,-.03
Accuracy in motor reaction (H2b)	Intervention completion score[a] (%)	83.3	19.3	69.1	22.9	167.5		.027[b]	0.33	
Speed in performance (H3)	Time on intervention tasks (z-score)	0.09	0.46	0.65	0.53	3.77	40	.001[b]	1.14	.26,.87
	Total time on scenario[a] (s)	619	172	714	180	348		.067	0.27	
Cognitive load (H4)	NASA-TLX score	42.88	12.66	52.94	16.18	2.33	40	.025[b]	0.70	1.35,18.77
	Average fixation duration (ms)	229.6	45.0	223.0	55.5	−0.41	36	.682	−0.13	−38.97,25.80
	Fixation frequency (s^{-1})	2.90	0.40	2.87	0.29	−0.30	36	.763	−0.10	-.26,.19
	Transition rate (s^{-1})	0.49	0.10	0.39	0.11	−3.09	39	.004[b]	−0.97	-.17,-.04

Note. [a]Due to non-normal distribution, Mann-Whitney *U* test was used. U-value and r were calculated instead of t-value and Cohen's d.
[b]$p < .05$.

(i.e., a low level of cognitive load). Average fixation duration and fixation frequency were calculated over the scenario. Transition was counted using all individual AOIs from the four AOI groups aforementioned. Then the per-second transition rate was calculated.

3. Results

All measures for each construct were compared between experts and novices by t-tests for independent samples. When the data is not normally distributed, Mann-Whitney *U* test was used instead. MANOVA was used for comparing multivariate variables. Table 2 provides an overview of the outcomes of the variables related to all constructs except visual selectivity that is specified separately in Table 3.

3.1. Systematicity in approach

Fig. 4 shows the distribution and the boxplot of HMM scores. The HMM score was significantly different between the two groups with a large effect size ($t(32) = -3.49$, $p = .001$; $d = -1.16$), indicating that experts adhered better to the ABCDE sequence than novices. There was no significant difference between groups in the length of the ABCDE sequences.

3.2. Accuracy in visual selection and motor reaction

Table 3 demonstrates an overview of outcomes of visual selectivity measures for each AOI category. A MANOVA was conducted for all three relative measures of visual selectivity (i.e., dwell time, fixation count, and fixation duration) for CDA. The MANOVA revealed a significant difference between experts and novices ($F(3,37) = 4.67$, $p = .007$). Further, separate t-tests on the all three variables showed significant difference: experts showed higher proportion of dwell time to total play time with a large effect size ($t(38) = -2.62$, $p = .012$; $d = -0.82$), higher ratio of fixation count to total fixation counts with a medium effect size ($U = 123$, $p = .023$, $r = 0.35$), and longer fixation duration with a large effect size (t(38) = −2.15, p = .038; d = −0.67) for CDA. There was no significant difference in other AOI groups, except for the experts' lower proportion of fixation count in IVA (t(38) = 2.07, p = .045; d = 0.65). The intervention completion score was significantly higher, with a medium effect size, for experts than for novices (U = 167.5, p = .027; r = 0.33).

3.3. Speed in performance

Experts showed faster unit-task reaction time with a large effect size ($t(40) = 3.77$, $p = .001$; $d = 1.14$). There was no significant difference in the total time on entire scenario performance ($U = 348$, $p = .067$; $r = 0.27$).

3.4. Cognitive load

A MANOVA was conducted for all four measures of cognitive load: NASA-TLX score, average fixation duration, fixation frequency, and transition rate. The MANOVA revealed a significant difference between experts and novices ($F(4,36) = 2.68$, $p = .047$) in cognitive load. However, separate t-tests showed diverged results. The NASA-TLX scores was lower for experts than novices, with a large effect size ($t(40) = 2.33$, $p = .025$; $d = 0.70$). There was no significant difference

Table 3
Visual selectivity for different AOI groups.

AOI group	Measure	Expert		Novice		*t* (*U*)	*df*	*p*	*d* (*r*)	95%CI
		M	*SD*	*M*	*SD*					
CDA	Dwell time (%)	0.32	0.10	0.24	0.09	−2.62	38	.012[b]	−0.82	-.14, −.02
	Fixation count[a] (%)	0.39	0.11	0.31	0.08	123		.023[b]		
	Fixation duration (ms)	420.0	148.7	328.6	123.2	−2.15	38	.038[b]	−0.67	−177.6, −5.35
NDA	Dwell time	0.09	0.03	0.07	0.03	−1.91	38	.063	−0.60	-.04, .00
	Fixation count	0.13	0.04	0.11	0.04	−1.84	37	.073	−0.58	-.05, −.00
	Fixation duration	174.3	73.0	133.0	69.1	−1.86	39	.070	−0.58	−86.19, 3.61
IVA	Dwell time	0.19	0.06	0.24	0.09	1.72	32	.095	0.55	-.01, .09
	Fixation count	0.27	0.08	0.31	0.07	2.07	38	.045[b]	0.65	.00, .10
	Fixation duration	269.4	90.7	225.6	80.7	−1.64	39	.110	−0.51	−98.99, 10.39
NA	Dwell time	0.10	0.01	0.09	0.04	−0.97	25	.339	−0.30	-.03, .01
	Fixation count	0.12	0.03	0.12	0.03	−0.21	38	.830	−0.07	.02, .02
	Fixation duration	222.9	60.6	224.7	71.5	0.09	37	.932	0.03	−40.24, 43.83

Note. Critical diagnosis area (CDA), non-critical diagnosis area (NDA), intervention area (IVA), and neutral area (NA). Each measure was calculated as relative values: dwell time divided by total play time, fixation count divided by total fixation counts, and fixation duration divided by fixation duration averaged over the scenario.
[a] Due to low normality, Mann-Whitney *U* test was used. U-value and r were calculated instead of t-value and Cohen's d.
[b] $p < .05$.

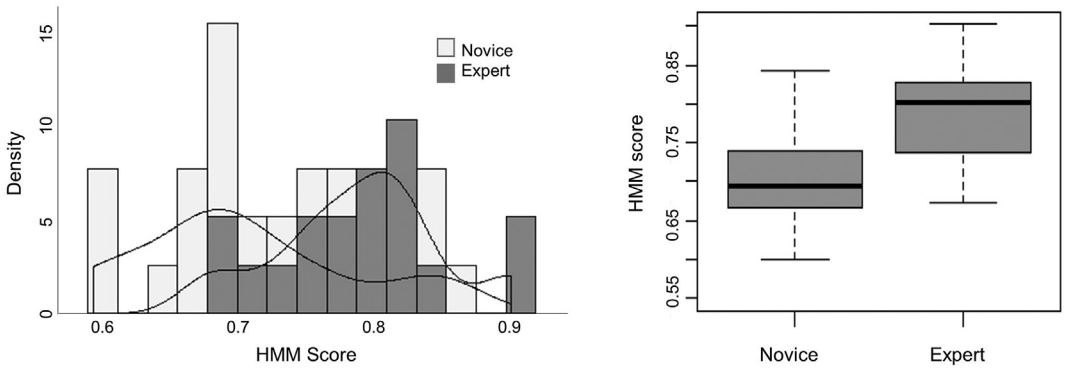

Fig. 4. Results of HMM score. The histogram on the left shows the distribution of HMM scores of experts (dark bars) and novices (light bars). The boxplot on the right depicts medians and quartiles of each group.

Table 4
Correlation between cognitive load measures.

Variable	1	2	3
1. NASA-TLX			
2. Average fixation duration	-.14 [-.43, .18]		
3. Fixation frequency	-.15 [-.44, .17]	.39[a] [.09, .62]	
4. Transition rate	-.39[a] [-.62, −.09]	.34[a] [.03, .58]	.64[b] [.41, .79]

Note. Values in square bracket indicate the 95% confidence interval for each correlation.
[a] $p < .05$.
[b] $p < .01$.

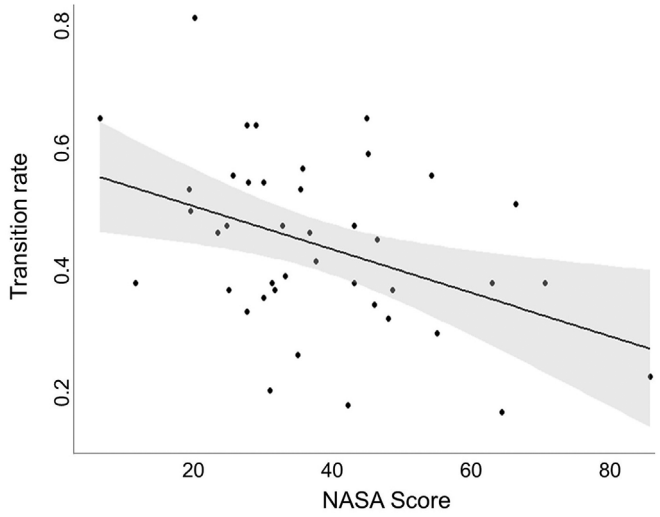

Fig. 5. Scatter plot of NASA-TLX scores and transition rate.

between experts and novices in average fixation duration ($t(36) = -0.41$, $p = .682$) and fixation count ratio ($t(36) = -0.30$, $p = .763$). Experts showed higher transition rate than novices, with a large effect size ($t(39) = -3.09$, $p = .004$; $d = -0.97$). Convergence between the NASA-TLX score and each eye-tracking measurement was examined via the Pearson correlation coefficient (Table 4). Transition rate displayed a negative correlation with the NASA-TLX scores ($r(39) = -0.39$, $p = .012$) (Fig. 5), while other eye-tracking measures were correlated between themselves.

4. Discussion

This study aimed to empirically determine whether the level of domain-specific prior knowledge (DSPK) affects performance in a computer-based simulation game (CBSG). In the introduction, we argued that, to assess this performance, certain constructs should be developed by taking theories of complex-skill acquisition as a starting point. We suggest four theoretical aspects of problem-solving performance to represent the level of DSPK, and defined indicators of these aspects based on a skill hierarchy, which resulted in four hypotheses. To confirm these hypotheses, game-logs and eye-tracking data were collected and analyzed, using the methods corresponding to each aspect.

Hypothesis 1 stated that participants with high DSPK (i.e., experts) show higher systematicity in their approach during performance in a CBSG than participants with low DSPK (i.e., novices). The results of this study support this hypothesis. Systematicity for the AbcdeSIM task environment was defined as a high level of adherence of the ABCDE sequence at unit-task level, while flexibly adjusting task performance at physical level. According to the result from the HMM, the experts showed a higher level of adherence than novices to the ABCDE sequence at a hidden level. Additionally, the length of the ABCDE sequences at the physical level did not show significant difference between experts and novices. This implies that the important difference between experts and novices resides in the inner structure of the task performance, rather than the amount of physical action.

Hypothesis 2 concerns the accuracy in applying cognitive rules, stating that experts show the availability of more accurate cognitive rules. We decomposed the cognitive rules into two parts specific to CBSG environments: visual input of information from the environment (i.e., *if* part) and motor reactions to the environment (i.e., *then* part). Therefore, the hypothesis consisted of two sub-hypotheses: Experts show higher accuracy in both visual selection (H2a) and motor reaction (H2b) than novices.

H2a was confirmed, showing that experts have higher accuracy in visual selection of areas with critical information. This construct was operationalized as the ratio of allocation of visual attentional resources to critical diagnosis areas (CDA). All three eye-tracking metrics that were used (i.e. dwell time, fixation count, and fixation duration) indicated a higher allocation to CDA for experts. We observed no significant difference in other AOI groups (i.e., NDA, IVA, and NA), except the fixation count in IVA. Interestingly, for the areas with intervention functions (IVA), experts showed lower fixation counts compared to novices. Regarding that high number of fixation counts indicate frequent reference to the stimulus (Orquin & Loose, 2013), this result suggests that novices search more frequently for what to execute (i.e., intervention) in the absence of an accurate diagnosis. This also can be interpreted as novices using weak methods such as *general search* (Gick, 1986) and *working backward* (Larkin, McDermott, Simon, & Simon, 1980). As a result of critical information collected via effective

information gathering strategies, experts reacted to the environment more accurately. They achieved higher intervention scores, which supports H2b. Thus, Hypothesis 2 was largely confirmed: Compared to novices, experts allocate more visual attentional resources to critical information, followed by more appropriate motor reactions.

With regard to H3 which concerns speed in performance, the results confirmed H3 by demonstrating higher speed in performing specific unit-tasks that are most essential for the designated scenario. On the other hand, one might note as well that the total play time on the entire scenario showed no significant difference between experts and novices. Although experts perform tasks faster than novices in general, we assume that the time on the entire task is not an applicable indicator of expertise in certain tasks. In this study, experts seem to complete the essential interventions faster, then repeatedly perform *reassessment* (i.e., monitoring the effect of applying the ABCDE procedure and controlling the process), resulting in a similar length of overall performance time between experts and novices. The reassessment is one of the essential parts of the ABCDE method, trained throughout the traineeship of emergency medicine, which is often overlooked by novices. We suggest that the use of time on entire task as an indicator of expertise should be considered carefully through analyzing given tasks.

Lastly, Hypothesis 4 pertained to the lower level of cognitive load for experts compared to novices. This was supported in that experts reported lower cognitive load in a subjective rating scale (NASA-TLX), which was correlated with high transition rate. While subjective rating scales are a well validated measure of cognitive load, the interpretation of transition rates has been inconsistent in the literature. A robust transition rate can be related with optimal use of working memory (Epelboim & Suppes, 2001; Miall & Tchalenko, 2001) or better integration between different information sources (Bartels & Marshall, 2006; Johnson & Mayer, 2012; Schmidt-Weigand, Kohnert, & Glowalla, 2010), which is on the same line with our interpretation. On the other hand, a high transition rate could also be connected to difficulties in integrating information sources (Holsanova, Holmberg, & Holmqvist, 2009), inefficient visual problem-solving strategies (Van Meeuwen et al., 2014), or extraneous cognitive load caused by seductive details in multimedia learning (Korbach, Brünken, & Park, 2017).

We presume that these different interpretations stem from differences in characteristics of visual stimuli which are highly task-dependent. When a task presents static stimuli with fixed information (e.g., static texts or figures), shifting eye-gazes between AOIs might indicate a stagnation within the same information. In this case, AOIs with information that has already been processed become irrelevant areas that does not require revisits (Van Meeuwen et al., 2014). On the other hand, when a task provides dynamically changing pieces of information, shifts between AOIs signifies rather a different kind of process, a vigorous progress in gathering new information. Especially in medical simulations, monitoring patients' physiological changes and reacting upon them constantly (i.e., reassessment) is a major part in problem-solving, which can be facilitated by optimal use of working memory. Our result accords with the results of a previous study in ultrasound simulation (Aldekhyl, Cavalcanti, & Naismith, 2018), which also used a medical simulation task. Furthermore, this dynamic aspect of medical simulation tasks seems to reduce the sensitivity of fixation duration and fixation count during the overall task performance in measuring cognitive load, due to the fluctuation of these measures. Further research is needed on using eye-tracking to measure cognitive load in different task environments.

The results of this study have several implications for indicator development in CBSGs. First of all, multiple aspects of performance should be considered as a whole within an integrated theoretical framework when determining constructs to assess performance. Researchers in education have argued that a well-designed performance assessment should combine all aspects of performance in a global manner, rather than using a simple checklist (Cunnington, Neville, & Norman, 1996; Dankbaar et al., 2014). We suggest that the same principle should be applied to the performance assessment in CBSGs. This is the most important reason to use complex-skill acquisition theories as a driving force, because it facilitates considering different aspects of tasks (e.g., non-routine and routine) in an integrated theoretical framework. Secondly, since constructs are abstract and conceptualized broadly, they should be operationalized to concrete indicators that are fully designated to a specific task. This should be done through a task analysis in consultation with domain experts, also driven by educational theories. For instance, systematicity in approach is one of the broad concepts we explored in this study, which was problematic to operationalize. Thus, we robustly applied theories in complex-skill acquisition and the relevant domain (i.e., the ABCDE method), to define the indicator of systematicity. Thirdly, this study opens the potential of combining eye-tracking data with game-log to quantify performance in CBSGs. Since most of CBSGs depend on visual stimuli, eye-tracking can be an important source to obtain a complete account of certain aspects of performance. In this study, we have found two indicators from eye-tracking (i.e., visual selection and cognitive load) that can possibly be used in CBSGs. Future research is needed to apply these findings to assessment and the development of support in CBSGs (e.g., student modelling to adaptively support individuals).

The insights from this study can help educators to assess students' performance in CBSGs and provide scaffolding to students with a low level of DSPK. For instance, they might focus on the different aspects in students' performance, then adjust the level of scaffolding to enhance each aspect. When the systematicity in approach is not high enough, instructors might stimulate the student to construct domain-specific knowledge and strategies (e.g., advise them to consult learning resources with relevant information). When a student concentrates on reactions only without sufficient information gathering, they can guide the student to pay more attention to information gathering as a sound foundation for taking actions in the game. Additionally, when the student's cognitive load is high, extra support could be given to manage the load. This can be done either by reducing the cognitive load itself (e.g., providing pauses during the game or presenting less complex scenarios), or facilitate self-regulation of students to manage their own cognitive load (Sweller et al., 2019).

Several limitations of this study need to be mentioned. Firstly, our findings might not be generalizable to other CBSG environments since the indicators were specialized for a specific task. Future research should follow to examine how our methods can be applied in other CBSG environments. Secondly, the participants in the expert group were composed of residents, rather than medical doctors. In this study, we selected medical students as novices and residents as experts, in order to form comparable groups. Although it led to a better controlled experimental setup, including a wider range of expertise levels could have yielded more informative results. Thirdly, although one could well argue that eye-hand coordination in performing cognitive rules is another aspect of performance, it was not explored in this study. We rather analyzed visual selection and motor reaction separately based on our assumption that those two cannot be matched one-to-one in a dynamic environment of a CBSG. However, investigating eye-hand coordination as an aspect of performance via non-linear analysis should be an intriguing topic for future study.

In conclusion, this study has demonstrated the development of performance assessment that can be used in a highly dynamic game environment. This was accomplished by starting from theories of complex-skill acquisition, identifying constructs for assessment and valid indicators. We believe empirical investigation for reliable indicators in CBSGs can be seen as a problem-solving process itself. As in problem-solving by experts, the research should be driven by a certain knowledge structure (e.g., educational theories) to avoid an inefficient process and suboptimal solutions. Educational theories and empirical experiment are in a reciprocal relationship, where one cannot stand alone without the other.

Acknowledgements

This work was supported by the Netherlands Organization for Scientific Research (NWO) [grant numbers 055.16.117]. The authors thank Tjitske Faber and Géraldine Sellenraad for providing medical expertise and anonymous reviewers for their insightful comments.

Appendix. The HMM score computation

We start computing the HMM score by first extracting the ABCDE phases of the subsequent observed actions from the log file. Next, a HMM is fitted to this sequence using an EM algorithm, as provided by the R package hmm.discnp (Rabiner, 1989; Turner, 2018). The HMM is set to have 5 inner states (actual phases) and 5 emission values (observed actions). Since fitting an HMM using a single observed sequence is strongly dependent on the starting condition, we initialize the HMM with a transition matrix with most probability mass concentrated on the diagonal and upper co-diagonal, and an emission probability matrix with most probability mass concentrated on the diagonal.

The resulting probability structure after fitting to the observed sequence contains information on the adherence to the ABCDE order. In the transition probability matrix, the total probability on “forbidden” transitions (e.g., jump from A to E) show how much is deviated from the order. The probability on “forbidden” emissions (e.g., an action for B in phase D) in the emission probability matrix shows how often actions are taken in a wrong phase. From this obtained probability structure, we compute a score: the total probability on legal transitions plus the total probability on legal emissions, divided by 2. This score ranges between 0 and 1.

Consequently, the HMM score increases when a performer keeps to the ABCDE phases in order, while the score decreases when the performance deviates from the order. For instance, in case of an ideal performer, the hidden sequence follows the ABCDE phases in a complete order (e.g., A-A-A-A-A-B-B-B-C-C-C-C-D-D-D-D-D-E-E-E-E-E). The HMM score for this example is 1.0. In the case of a less ideal performer, the sequence may deviate from the complete order (e.g., A-A-A-B-E-C-C-E-D-B-C-A-C-B-D-C-E-C-D-E-D-A-E-E), signifying that this performer jumped around the phases using less ideal rules. The HMM score for this example is 0.792.

References

Aldekhyl, S., Cavalcanti, R. B., & Naismith, L. M. (2018). Cognitive load predicts point-of-care ultrasound simulator performance. *Perspectives on medical education,* 1–10.

Bartels, M., & Marshall, S. P. (2006). Eye tracking insights into cognitive modeling. *Paper presented at the Proceedings of the 2006 symposium on Eye tracking research & applications*.

Bartlett, F. C. (1995). *Remembering: A study in experimental and social psychology*, *Vol. 14*. Cambridge University Press.

Baum, L. E., Petrie, T., Soules, G., & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics, 41*(1), 164–171.

Clark, D. B., Martinez-Garza, M. M., Biswas, G., Luecht, R. M., & Sengupta, P. (2012). Driving assessment of students' explanations in game dialog using computer-adaptive testing and hidden Markov Modeling. In D. Ifenthaler, D. Eseryel, & X. Ge (Eds.). *Assessment in game-based learning* (pp. 173–199). New York, NY: Springer.

Cunnington, J., Neville, A., & Norman, G. (1996). The risks of thoroughness: Reliability and validity of global ratings and checklists in an OSCE. *Advances in Health Sciences Education, 1*(3), 227–233.

Dankbaar, M. E., Alsma, J., Jansen, E. E., van Merrienboer, J. J., van Saase, J. L., & Schuit, S. C. (2016). An experimental study on the effects of a simulation game on students' clinical cognitive skills and motivation. *Advances in Health Sciences Education, 21*(3), 505–521.

Dankbaar, M. E., Stegers-Jager, K. M., Baarveld, F., van Merrienboer, J. J., Norman, G. R., Rutten, F. L., ... Schuit, S. C. (2014). Assessing the assessment in emergency care training. *PLoS One, 9*(12), e114663. Retrieved from https://doi.org/10.1371/journal.pone.0114663.

De Freitas, S. (2006). Using games and simulations for supporting learning. *Learning, Media and Technology, 31*(4), 343–358 10.1080/17439880601021967.

Donovan, T., & Litchfield, D. (2013). Looking for cancer: Expertise related differences in searching and decision making. *Applied Cognitive Psychology, 27*(1), 43–49.

Epelboim, J., & Suppes, P. (2001). A model of eye movements and visual working memory during problem solving in geometry. *Vision Research, 41*(12), 1561–1574.

Erasmus University Medical Center and VirtualMedSchool (2012). *AbcdeSIM*. Rotterdam: VirtualMedSchool. Retrieved from https://virtualmedschool.com/.

Ericsson, K. A., Hoffman, R. R., Kozbelt, A., & Williams, A. M. (2018). *The Cambridge handbook of expertise and expert performance*. Cambridge University Press.

Gagne, R. M. (1968). Learning hierarchies. *Educational Psychologist, 6*(1), 1–9.

Gibson, J. J. (2014). *The ecological approach to visual perception: Classic edition*. New York, NY: Psychology Press.

Gick, M. L. (1986). Problem-solving strategies. *Educational Psychologist, 21*(1–2), 99–120.

Grigg, S. J., Garrett, S. K., & Benson, L. C. (2012). Using the NASA-TLX to assess first year engineering problem difficulty. *IIE Annual Conference. Proceedings* (pp. 1). Institute of Industrial and Systems Engineers (IISE).

Haider, H., & Frensch, P. A. (1999). Eye movement during skill acquisition: More evidence for the information-reduction hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*(1), 172.

Hart, S. G., & Staveland, L. E. (1988). *Development of NASA-TLX (task load Index): Results of empirical and theoretical research. Advances in psychology*, *Vol. 52*, Elsevier139–183.

Hernández-Lara, A. B., Perera-Lluna, A., & Serradell-López, E. (2019). Applying learning analytics to students' interaction in business simulation games. The usefulness of learning analytics to know what students really learn. *Computers in Human Behavior, 92*, 600–612.

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: OUP.

Holsanova, J., Holmberg, N., & Holmqvist, K. (2009). Reading information graphics: The role of spatial contiguity and dual attentional guidance. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition, 23*(9), 1215–1226.

Hou, H.-T. (2015). Integrating cluster and sequential analysis to explore learners' flow and behavioral patterns in a simulation game with situated-learning context for science courses: A video-based process exploration. *Computers in Human Behavior, 48*, 424–435.

Johnson, C. I., & Mayer, R. E. (2012). An eye movement analysis of the spatial contiguity effect in multimedia learning. *Journal of Experimental Psychology: Applied, 18*(2), 178.

Kang, J., Liu, M., & Qu, W. (2017). Using gameplay data to examine learning behavior patterns in a serious game. *Computers in Human Behavior, 72*, 757–770.

Kiili, K. (2005). Digital game-based learning: Towards an experiential gaming model. *The Internet and Higher Education, 8*(1), 13–24.

Koivisto, J.-M., Niemi, H., Multisilta, J., & Eriksson, E. (2017). Nursing students' experiential learning processes using an online 3D simulation game. *Education and Information Technologies, 22*(1), 383–398.

Korbach, A., Brünken, R., & Park, B. (2016). Learner characteristics and information processing in multimedia learning: A moderated mediation of the seductive details effect. *Learning and Individual Differences, 51*, 59–68.

Korbach, A., Brünken, R., & Park, B. (2017). Measurement of cognitive load in multimedia learning: A comparison of different objective measures. *Instructional Science, 45*(4), 515–536.

Larkin, J., McDermott, J., Simon, D. P., & Simon, H. A. (1980). Expert and novice performance in solving physics problems. *Science, 208*(4450), 1335–1342.

Lee, F. J., & Anderson, J. R. (2001). Does learning a complex task have to be complex?: A study in learning decomposition. *Cognitive Psychology, 42*(3), 267–316.

Liversedge, S., Gilchrist, I., & Everling, S. (2011). *The Oxford handbook of eye movements*. Oxford University Press.

Loh, C. S., Sheng, Y., & Ifenthaler, D. (2015). Serious games analytics: Theoretical framework. *Serious games analytics* (pp. 3–29). Springer.

Manning, D., Ethell, S., Donovan, T., & Crawford, T. (2006). How do radiologists do it? The influence of experience and training on searching for chest nodules. *Radiography, 12*(2), 134–142.

McLaughlin, L., Bond, R., Hughes, C., McConnell, J., & McFadden, S. (2017). Computing eye gaze metrics for the automatic assessment of radiographer performance during X-ray image interpretation. *International Journal of Medical Informatics, 105*, 11–21.

Miall, R. C., & Tchalenko, J. (2001). A painter's eye movements: A study of eye and hand movement during portrait drawing. *Leonardo, 34*(1), 35–40.

Newell, A., & Simon, H. A. (1972). *Human problem solving*, *Vol. 104*Englewood Cliffs, NJ: Prentice-Hall No. 9.

Orquin, J. L., & Holmqvist, K. (2018). Threats to the validity of eye-movement research in psychology. *Behavior Research Methods, 50*(4), 1645–1656.

Orquin, J. L., & Loose, S. M. (2013). Attention and choice: A review on eye movements in decision making. *Acta Psychologica, 144*(1), 190–206.

Palmeri, J. T. (1999). Theories of automaticity and the power law of practice. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 543–551. https://doi.org/10.1037/0278-7393.25.2.543.

Park, B., Knörzer, L., Plass, J. L., & Brünken, R. (2015). Emotional design and positive emotions in multimedia learning: An eyetracking study on the use of anthropomorphisms. *Computers & Education, 86*, 30–42.

R Development Core Team (2010). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Retrieved from http://www.R-project.org.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE, 77*(2), 257–286.

Reingold, E. M., & Sheridan, H. (2011). *Eye movements and visual expertise in chess and medicine*. Oxford handbook on eye movements.

Schmidt-Weigand, F., Kohnert, A., & Glowalla, U. (2010). A closer look at split visual attention in system-and self-paced instruction in multimedia learning. *Learning and Instruction, 20*(2), 100–110.

Shamo, M. K., Dror, R., & Degani, A. (1999). A multi-dimensional evaluation methodology for new cockpit systems. *Paper presented at the Proceedings of the Tenth International symposium on aviation Psychology*.

Squire, K. (2008). Open-ended video games: A model for developing learning for the

interactive age. *The ecology of games: Connecting youth, games, and learning,* 167–198.
Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*(2), 257–285.
Sweller, J., Clark, R., & Kirschner, P. (2010). Teaching general problem-solving skills is not a substitute for, or a viable addition to, teaching mathematics. *Notices of the American Mathematical Society, 57*(10), 1303–1304.
Sweller, J., van Merriënboer, J. J., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review,* 1–32.
Turner, R. (2018). *Package 'hmm.discnp'.* Retrieved from https://cran.r-project.org/web/packages/hmm.discnp.
Van Meeuwen, L. W., Jarodzka, H., Brand-Gruwel, S., Kirschner, P. A., de Bock, J. J., & van Merriënboer, J. J. (2014). Identification of effective visual problem solving strategies in a complex visual domain. *Learning and Instruction, 32*, 10–21.
Van Merriënboer, J. J. (2013). Perspectives on problem solving and instruction. *Computers & Education, 64*, 153–160.
Van Merriënboer, J. J., & Kirschner, P. A. (2018). *Ten steps to complex learning: A systematic approach to four-component instructional design.* Routledge.
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.
Van Orden, K. F., Nugent, W., La Fleur, B., & Moncho, S. (1998). *Assessment of variable coded symbology using visual search performance and eye fixation measures (Report No. 99-4)*San Diego, CA: Naval Health Research Center.
Wallner, G., & Kriglstein, S. (2013). Visualization-based analysis of gameplay data–a review of literature. *Entertainment Computing, 4*(3), 143–155.
Warm, J. S., Matthews, G., & Finomore, V. S., Jr. (2017). Vigilance, workload, and stress. In J. L. Szalma, & P. A. A. Hancock (Eds.). *Performance under stress* (pp. 131–158). London, England: CRC Press.
Weinger, M. B., Vredenburgh, A. G., Schumann, C. M., Macario, A., Williams, K. J., Kalsher, M. J., ... Kim, A. (2000). Quantitative description of the workload associated with airway management procedures. *Journal of Clinical Anesthesia, 12*(4), 273–282.
Zelinsky, G. J., Rao, R. P. N., Hayhoe, M. M., & Ballard, D. H. (1997). Eye movements reveal the spatiotemporal dynamics of visual search. *Psychological Science, 8*(6), 448–453.
Zhou, M., Xu, Y., Nesbit, J. C., & Winne, P. H. (2010). Sequential pattern analysis of learning logs: Methodology and applications. *Handbook of educational data mining, 107*.