A new way to look at simulation-based assessment: the relationship between gaze-tracking and exam performance

Adam Szulewski, MD, MHPE*; Rylan Egan, PhD†; Andreas Gegenfurtner, PhD‡; Daniel Howes, MD*; Gerhard Dashi, BSc§; Nick C. J. McGraw, BSc¶; Andrew K. Hall, MD, MMed*; Damon Dagnone, MD, MMed*; Jeroen J. G. van Merrienboer, PhD‖

CLINICIAN'S CAPSULE

What is known about this topic?
Gathering visual information effectively is an important task of a physician leader when managing a resuscitation case.
What did this study ask?
Are there particular visual information-gathering patterns associated with performance in simulated resuscitation scenarios?
What did this study find?
Certain visual patterns (e.g., focusing on case-specific clinically relevant stimuli) are associated with better performance in a simulated resuscitation setting.
Why does this study matter to clinicians?
The ability to characterize physician visual patterns across a competence continuum has implications for trainee assessment and medical education.

ABSTRACT

Objective: A key task of the team leader in a medical emergency is effective information gathering. Studying information gathering patterns is readily accomplished with the use of gaze-tracking glasses. This technology was used to generate hypotheses about the relationship between performance scores and expert-hypothesized visual areas of interest in residents across scenarios in simulated medical resuscitation examinations.

Methods: Emergency medicine residents wore gaze-tracking glasses during two simulation-based examinations (n = 29 and 13 respectively). Blinded experts assessed video-recorded performances using a simulation performance assessment tool that has validity evidence in this context. The relationships between gaze patterns and performance scores were analyzed and potential hypotheses generated. Four scenarios were assessed in this study: diabetic ketoacidosis, bradycardia secondary to beta-blocker overdose, ruptured abdominal aortic aneurysm and metabolic acidosis caused by antifreeze ingestion.

Results: Specific gaze patterns were correlated with objective performance. High performers were more likely to fixate on task-relevant stimuli and appropriately ignore task-irrelevant stimuli compared with lower performers. For example, shorter latency to fixation on the vital signs in a case of diabetic ketoacidosis was positively correlated with performance ($r = 0.70$, $p < 0.05$). Conversely, total time spent fixating on lab values in a case of ruptured abdominal aortic aneurysm was negatively correlated with performance ($r = -0.50$, $p < 0.05$).

Conclusions: There are differences between the visual patterns of high and low-performing residents. These findings may allow for better characterization of expertise development in resuscitation medicine and provide a framework for future study of visual behaviours in resuscitation cases.

RÉSUMÉ

Objectif: Une des tâches importantes du chef d'équipe en médecine d'urgence est la collecte efficace de renseignements, et le port de lunettes de monitorage oculaire permet d'étudier rapidement et facilement les différents modes de collecte visuelle de renseignements. Les chercheurs ont donc eu recours à cette technique pour émettre des hypothèses selon lesquelles il existerait une relation entre la performance des résidents et la saisie de champs visuels d'intérêt présumés tels par des experts, dans des scénarios d'examen de réanimation axés sur la simulation, en médecine.

Méthode: Des résidents en médecine d'urgence ont porté des lunettes de monitorage oculaire au cours de deux séances d'examen par simulation (n = 29 et n = 13, respectivement).

From the *Department of Emergency Medicine, Queen's University, Kingston, ON; †Office of Health Sciences Education, Queen's University, Kingston, ON; ‡Researcher, Technische Hochschule Deggendorf, Deggendorf, Germany; §School of Medicine, Queen's University, Queen's University, Kingston, ON; ¶Faculty of Medicine, University of Toronto, Toronto, ON; and ‖Department of Educational Development and Research, Faculty of Health, Medicine and Life Sciences, Maastricht University, Maastricht, The Netherlands.

Correspondence to: Dr. Adam Szulewski; Assistant Professor, Department of Emergency Medicine, Queen's University, Kingston General Hospital, Victory 3, 76 Stuart Street, Kingston, ON K7L 2V7; Email: aszulewski@qmed.ca

CJEM 2019;21(1):129-137 DOI 10.1017/cem.2018.391

Des experts tenus dans l'ignorance ont évalué le comportement des résidents enregistré sur vidéo, à l'aide d'un outil d'évaluation de la performance axée sur la simulation, dont la validité avait déjà fait ses preuves dans le contexte. Il y a eu analyse de relations entre la collecte visuelle de renseignements et la performance ainsi qu'émission d'hypothèses. Quatre scénarios ont été évalués dans l'étude : l'acidocétose diabétique, une bradycardie consécutive à un surdosage de bêta-bloquants, une rupture d'anévrisme de l'aorte abdominale et une acidose métabolique causée par la consommation d'antigel.

Résultats: Des corrélations ont été établies entre certaines manières de recueillir visuellement les renseignements et la performance objective. Ainsi, ceux qui ont le mieux réussi avaient davantage tendance à se concentrer sur les stimulus pertinents, relatifs aux tâches et à ne pas tenir compte, à juste titre, des stimulus non pertinents comparativement à ceux qui ont moins bien réussi. Par exemple, un temps d'attente raccourci dans la collecte visuelle de renseignements sur les signes vitaux dans le cas de l'acidocétose diabétique s'est révélé en corrélation positive avec la performance ($r = 0,70$; $p < 0,05$), tandis que le temps total écoulé à glaner des renseignements sur les résultats d'examens de laboratoire dans le cas de la rupture d'anévrisme de l'aorte abdominale était en corrélation négative avec la performance ($r = -0,50$; $p < 0,05$).

Conclusion: Il existe donc des différences entre la manière de glaner visuellement les renseignements et la performance, bonne ou mauvaise, des résidents. Les résultats pourraient permettre une meilleure caractérisation de l'acquisition de la compétence en matière de médecine de réanimation, et fournir un cadre de travail en vue d'études futures sur les comportements visuels dans les cas de réanimation.

Keywords: assessment, emergency medicine, gaze-tracking, medical education, simulation

INTRODUCTION

Successful management of a medical emergency demands effective crisis resource management (CRM) skills.[1,2] Physicians skilled in CRM are usually easily identifiable by their peers, but describing specific behaviours that make them successful in crisis settings is difficult, even for physicians themselves.[3] A reason is that high-level CRM skills are automatized over many years as a result of deliberate practice and become second nature.[4] This tacit knowledge is subsequently difficult to describe explicitly and teach.[5,6]

Tacit knowledge poses particular challenges to medical educators who are tasked with preparing and assessing learners in routine and emergent medical cases. It follows that the more that is known about expertise development in the management of medical emergencies, such as resuscitation and CRM, the more likely it will be that both can be taught and assessed effectively and efficiently. Moreover, by triangulating novel objective measures of learner performance with traditional assessment, educators can calibrate both modalities and increase the validity of both subjective and objective means of trainee assessment. This would improve the granularity of resuscitation assessment, with decreased risks to patients, a need that will only increase as medical schools shift to competency-based medical education (CBME).

One commonly used approach to teaching CRM skills well suited to integrating CBME is simulation-based training.[7] High-fidelity medical simulation-based training improves learning by providing a forum for feedback, integration of curriculum, and an opportunity for repetitive practice for routine cases, as well as variability of practice for non-routine cases.[8-10] Formative and summative assessments in simulation are also gaining popularity,[11] especially in postgraduate medical education in which assessment tools are being developed and optimized.[12-14]

It is difficult to glean a comprehensive view of expertise development in resuscitation medicine using traditional observational research methods. A novel method of studying resuscitation medicine and CRM expertise is with the use of mobile gaze-tracking technology. Gaze-tracking data can provide new insights into clinician behaviour and performance and may lead to improved patient safety practices.[15] Mobile gaze-tracking glasses are worn like a pair of eyeglasses that record both a video of a participant's first-person point-of-view and track the participant's eye movements, superimposing a gaze indicator on the first-person video, revealing where the participant is looking in real time. Based on data that show what a physician "sees," inferences can be made on the differences between novices and experts with respect to their visual attention, situational awareness, and ability to manage competing interests.[16]

Visual expertise research has shown that there are measurable visual differences between novices and experts in numerous domains that can be quantified using gaze-tracking technology.[17,18] Experts can selectively ignore irrelevant information, recognize patterns

https://doi.org/10.1017/cem.2018.391 Published online by Cambridge University Press

rapidly, and select appropriate diagnostic schemata to fit what they see.[19] With practice, individuals become better, more efficient interpreters of their surroundings. They are better able to prioritize what is important while simultaneously deprioritizing what is not. Analyzing eye movements in a non-medical field, Haider and Frensch[20] termed this phenomenon the "information-reduction hypothesis," and Kok et al.[21] hypothesized that there may be parallels to be drawn in medicine.

Gaze-tracking technology has also been used to determine the feasibility of studying physician behaviours during simulated medical emergencies, and it has been shown that this technique is practical and useful in this setting.[3] However, little is known about whether physicians' expertise in the complex environment of resuscitation medicine can be accurately assessed by studying their visual patterns.

Building on this research, this study explored the information-gathering techniques of residents by analyzing their initial visual fixation patterns in a simulated resuscitation environment. Specifically, we were interested in uncovering particular gaze patterns used by physicians that are associated with better exam performance. The objectives of this study were to: 1) examine the relationship between initial visual patterns and performance scores; and 2) explore if these patterns varied in different simulation scenarios.

METHODS

Study design and participants

A convenience sample of emergency medicine (EM) residents at one training site were invited to participate in two exams, each comprised of two simulated scenarios.

The first two scenarios (S1, diabetic ketoacidosis; and S2, bradycardia secondary to beta-blocker overdose) were conducted in February 2014; the second two scenarios (S3, ruptured abdominal aortic aneurysm [AAA]; and S4, metabolic acidosis caused by antifreeze ingestion) were conducted in August 2014. The scenarios and their associated assessment tools were developed through an iterative process that consisted of a template-based scenario development tool. At least three expert EM physicians reviewed each case for clinical fidelity. A previous study of EM residents, using similar scenarios developed in this manner, demonstrated excellent reliability (as determined through a G-study analysis).[12] The study took place during regularly scheduled biannual simulation-based objective structured clinical exams (OSCEs). Participants had previously participated in similar OSCEs and were familiar with the assessment environment and format; however, the scenarios used in this study were novel to each participant. Participants did not receive any incentives to participate. The study was approved by the Research Ethics Board at Queen's University (EMED-162-11).

Experimental design

Residents were fitted and calibrated with a mobile gaze-tracking device (Tobii Glasses Eye Tracker, Danderyd, Sweden). After reading a case stem, they entered the high-fidelity simulation lab as the leader of a team comprised of a registered nurse (RN) actor and respiratory therapist (RT) actor. Simulation rooms were organized to mimic the resuscitation bay in the emergency department (ED). The RN and RT began the scenarios in the same locations in the room; however, they were free to move around the room based on instructions given by the resident. Participants completed two 10-minute scenarios. Audio and video data from three (non-gaze-tracking) cameras were also recorded (Kb Port, Allison Park, PA). These videos were subsequently used by the blinded external reviewers to score participant performance.

Analysis

The gaze-tracking device recorded first-person audio and video data and gathered information on pupil position and glint location at a rate of 30 Hz. Using these data, a dynamic gaze indicator was superimposed on the first-person video by computer software (Tobii Studio Pro). This video was used in the analysis to compute the gaze-tracking variables and fixation areas listed in Table 1. Figures 1 and 2 show static representations of the data generated by the software. In this study, fixations referred to instances when the gaze indicator stopped scanning the environment and landed on an area of interest. The frequency and duration of the residents' gazes were scored independently by two trained individuals who manually analyzed each first-person video. Human raters were used in this study to count fixations, as computer technology that can

https://doi.org/10.1017/cem.2018.391 Published online by Cambridge University Press

Table 1. Measured gaze-tracking fixation areas

Time in first 60 seconds	Number of visual fixations in the first 60 seconds on the
To the first vital signs fixation	Patient
In silence	RN
Fixating on vital signs	RT
Fixating on lab values	Vitals
Fixating on RN	Medication list
Fixating on ECG	

ECG = electrocardiogram; RN = registered nurse; RT = respiratory therapist.

Figure 1. A third year resident's gaze fixations within the first minute of a simulation. The numbers represent the order in which the resident looked at each area of interest and the size of the circle represents the relative time spent fixating on each of these points.

Figure 2. A third year resident's gaze fixations within the first minute of a simulation. The colours represent a heat map, where red represents more fixations for a longer duration, followed by yellow and green areas, which represent fewer fixations.

accurately triangulate eye-tracking data with three-dimensional positional data in a resuscitation room with dynamic areas of interest (where individuals and equipment move) is still under development. The dichotomous actions and visual propensities of the participants were captured (see Table 1), and Pearson correlation coefficients were used to determine whether associations existed with performance. These areas of interest were defined *a priori* based on their potential clinical relevance and input from local medical experts. We focused this analysis on the first 60 seconds of each scenario, as we were mostly interested in how residents' initial information-gathering patterns correlated with performance scores. Furthermore, our previous experience suggested that most relevant visual fixations occur early in simulated scenarios and that there is increased variability and noise in gaze patterns as the simulations progress.

Inter-rater reliability between the two individuals tabulating eye-tracking variables and the expert performance assessors were determined by calculating intraclass correlation coefficients (ICC). For each eye-tracking video, an ICC for aggregated fixation times (total fixation time and time until first vital signs check) and the net number of object fixations in the first 60 seconds were calculated. Two independent undergraduate reviewers used video time stamps and systematic video pausing to record the total gaze duration and frequencies. Reliability was calculated separately for S1 and S2, as well as S3 and S4, as only five participants were involved in all four scenarios.

Performance of the residents was assessed by external attending emergency physicians who were blinded to the participants' identity and level of training. The Queen's Simulation Assessment Tool (QSAT) was used for assessment by the physicians after their review of the non-gaze-tracking recordings. The raters had previous experience and training on using the QSAT in this context. The QSAT has been previously shown to discriminate simulation-based OSCE performance with a similar cohort of EM residents within a CRM context.[12] The QSAT (see Appendix 1) assessed residents on a five-point Likert scale from inferior to superior across four categories (primary assessment, diagnostic actions, therapeutic actions, and communication), as well as overall performance.

A factorial analysis of variance (ANOVA) was conducted to determine if score differentials were only a function of training level or if gaze-tracking data provided information on the level of experience that extended beyond the level of training. Grouping of residents by experience in S1/S2 and S3/S4 differed because of uneven numbers of participants across postgraduate years (PGY). Of note, CCFP (PGY-3)

https://doi.org/10.1017/cem.2018.391 Published online by Cambridge University Press

residents were considered independent of Fellow of the Royal College of Canada (FRCPC) PGY-3 residents.

RESULTS

Participants

As for participation, 13 and 29 residents took part in the S1/S2 and S3/S4 exams, respectively (five completed both exams). One participant declined to consent to gaze tracking (but still completed the exam). Because of poor data quality and technical errors, the results from three cases in the first OSCE and three cases in the second were excluded. Therefore, a total of 78 cases were subsequently analyzed. Participants were EM residents enrolled in the College of Family Physicians of Canada (CCFP) and the Royal College of Physicians and Surgeons of Canada (RCPSC) programs (see Table 2) at one EM training site (mean age 30.0 years, standard deviation [SD] 2.91, and 48% were female residents).

Gaze-tracking data

Appendix 2 outlines the average fixation data for the scenarios. An acceptable average ICC was found for the fixation time ratings between the two individuals who analyzed the gaze-tracking videos in S1 and S2 (ICC = 0.87, $p < 0.001$, 95% confidence interval [CI] 0.15–0.97); however, an unacceptable ICC was found for S3 and S4 (ICC = 0.39, $p < 0.001$, 95% CI −0.22 to 0.71). An acceptable average ICC was found for the net number of object fixations (first 60 seconds) for S1 and S2 (ICC = 0.67, $p < 0.032$, 95% CI −0.03 to 0.90]) and S3 and S4 (ICC = 0.82, $p < 0.001$, 95% CI 0.59–0.92). Fixation time ratings were thought to be unsatisfactory largely because of the difficulty of manually recording the time and number of gazes. The future development of automated tracking processes would be beneficial to the accuracy of tracking.

Table 2. Number of participants by level of training and simulated case for the analyzed data

Level of training	Diabetic ketoacidosis	Bradycardia secondary to beta-blocker overdose	Ruptured AAA	Antifreeze ingestion
PGY-1	0	0	5	5
PGY-2	1	2	6	7
PGY-3	2	2	2	2
PGY-4	1	1	3	3
PGY-5	3	3	1	1
CCFP (PGY-3)	4	4	10	10
Total*	11	12	27	28

AAA = abdominal aortic aneurysm; CCFP = College of Family Physicians of Canada; PGY = postgraduate year.
*A total of six cases were excluded because of poor data quality and technical errors.

Performance scores

The calculated ICC was very good between the two expert assessors in S1 (ICC = 0.85, $p < 0.01$, 95% CI 0.45–0.96) and S2 (ICC = 0.87, $p < 0.001$, 95% CI 0.02–0.97). Only one external blinded assessor reviewed S3 and S4. Assessors scored participants across four categories for a total score, in addition to an overall impression score. A two-way mixed ICC was calculated for absolute accuracy between the raters for the total and overall scores. The findings showed acceptable reliability between the total and overall scores for S1 (ICC = 0.933, $p < 0.001$, 95% CI 0.65–0.98]), S2 (ICC = 0.987, $p < 0.001$, 95% CI 0.95–0.99), S3 (ICC = 0.95, $p < 0.001$, 95% CI 0.86–0.98), and S4 (ICC = 0.87, $p < 0.001$, 95% CI 0.70–0.94). Because of the congruency between the total and overall scores and between assessors, an average was taken between the percent total and overall scores for each scenario to create a new average score. This final score was used for subsequent analyses (see Appendix 3). After checking for normality, correlations were computed to determine associations between the data points in Table 1 and simulation performance scores.

Relationship between the level of training and performance

In S1 and S2, residency years were separated into three categories: PGY-1 to PGY-4 (mean or M; M S1 = 0.57, S2 = 0.60), PGY-5, (M S1 = 0.90, S2 = 0.93), and CCFP (PGY-3) (M S1 = 0.58, S2 = 0.56). A significant difference was demonstrated regarding the main effects between groups for S1 (F statistic or F;F[2, 8] = 5.32, $p = 0.03$, $\beta = 0.67$) and S2 (F[2, 9] = 4.90, $p = 0.04$, $\beta = 0.65$). There were significant differences between PGY-5 and CCFP PGY-3 residents ($p = .034$) in S1. No significant differences were found among the groups in S2. In S3 and S4, residency years were

https://doi.org/10.1017/cem.2018.391 Published online by Cambridge University Press

separated into three categories: PGY-1 and PGY-2 (M S3 = 0.64, S4 = 0.59), PGY-3 to PGY-5 (M S3 = 0.75, S4 = 0.63), and CCFP (PGY-3) (M S3 = 0.66, S4 = 0.57). No detectable significant difference was demonstrated between residency groupings in S3 ($F[2, 25] = 1.65$, $p = 0.2$, $\beta = 0.32$) and S4 ($F[2, 25] = 0.14$, $p = 0.86$, $\beta = 0.07$). Because of differences in the number of participants and spread of the residents' level of training between S1/S2 and S3/S4, the groupings could not be consistent. Independent *t*-tests were used to determine the ordering effects across scenarios. No effects of ordering were found ($p > 0.4$). A sample size calculation was not performed, as we maximized the number of residents who were available to participate. Instead, we have provided statistical power calculations (β).

Relationship between gaze tracking and performance

The observed correlations between the gaze-tracking variables and average performance score are shown in Tables 3–5 (Please see Appendix 4 for the complete correlation tables). Significant correlations in S1, S3, and S4 indicated that visually fixating on particular people and objects was correlated with performance.

DISCUSSION

Residents' gaze-tracking patterns were found to be significantly correlated with objective performance in the simulation-based resuscitation examinations. Because of the number of participants and array of data, correlation was used to elucidate patterns and associations. Given that this was a pioneering study on the use of gaze analytics to establish patterns associated with performance, these results have not been put forward as definitive findings; rather, they lay the groundwork for future study in this novel area.

In S1 (unstable patient with diabetic ketoacidosis), performance was strongly and positively correlated with decreased latency in checking the patient's vital signs, a task-relevant stimulus (Pearson's r 0.70, $p < 0.05$). The number of times a participant looked at the RT was negatively correlated with performance (Pearson's r – 0.65, $p < 0.05$). In this case, the RT did not provide any useful clinical information and also did not have a predefined script. In addition, the patient did not require any advanced airway intervention and did not have any increased work during breathing; therefore, it is plausible that increased views were not only task irrelevant but also potentially associated with help seeking or other attempts at social validation.

In S2 (beta-blocker overdose), there were no statistically detectable correlations identified. This is likely a result of multiple factors. First, the scenario and corresponding assessment tool were unable to discriminate by the level of training. Anecdotally, it seems that there was malalignment between the scenario design features and expected trainee behaviours. For example, many senior trainees did not perform transcutaneous pacing and focused instead on applying antidotes and medical management immediately for beta-blocker toxicity. Retrospectively, this was not an unreasonable management decision. However, raters were informed that transcutaneous pacing was an important therapy for this patient and likely rated these performances with a lower score.

Table 3. Correlations between selected gaze indicators and the average score in S1 (diabetic ketoacidosis)

Variables	1	2	3
Average score	-		
Number of RT fixations (first 60 seconds)	–0.65*	-	
Time to the first fixation on vital signs	–0.70*	0.48	-

RT = respiratory therapist.
*$p < 0.05$.

Table 4. Correlations between selected gaze indicators and the average score in S3 (ruptured AAA)

Variables	1	2	3	4
Average score	-			
Total time fixating on lab values (first 60 seconds)	–0.50*	-		
Total ECG fixation time (first 60 seconds)	–0.39*	0.16	-	
Number of RN fixations (first 60 seconds)	–0.46*	0.54†	0.117	-

ECG = electrocardiogram; RN = registered nurse.
*$p < 0.05$, †$p < 0.01$.

Table 5. Correlations between selected gaze indicators and the average score in S4 (metabolic acidosis)

Variables	1	2
Average score	-	
Number of medication list fixations	–0.48*	-

*$p < 0.05$.

https://doi.org/10.1017/cem.2018.391 Published online by Cambridge University Press

In S3 (patient with a ruptured AAA), we found that visual fixations on task-irrelevant stimuli such as the laboratory values, electrocardiogram (ECG), and nurse were all strongly and negatively correlated with objective performance (Pearson's r −0.50, −0.39, and −0.46, respectively; $p < 0.05$). A physician who made a diagnosis of a ruptured AAA should be primarily focused (after initial stabilization) on transporting the patient to the operating room for life-saving surgery. On video review, it was observed that higher performers expedited patient transfer by concentrating on contacting the vascular surgeon, calling the operating room charge nurse, and ensuring that the patient was on portable monitors and ready for rapid transfer. We did not tabulate these areas of interest in the data collection, as these tasks were not predicted to be relevant *a priori*.

In S4 (patient with metabolic acidosis caused by antifreeze ingestion), participants needed to provide primary resuscitation and appropriately manage a patient with an undifferentiated metabolic acidosis. The number of times a participant fixated on the medication list was correlated with poorer performance (Pearson's r −0.48, $p < 0.05$). These low-performing participants tended to assume that a medication overdose was the cause of the patient's condition and overlooked clues that pointed to ingestion of another toxic compound (antifreeze in this case). Those who performed better identified the appropriate agent and spent less time viewing the patient's medication list.

As in clinical practice, each scenario was different from the others and required medical learners to focus their attention on differentiating stimuli to be successful. Because these visual patterns differ from scenario to scenario, it would be difficult for a learner to feign superior visual behaviours without having the requisite knowledge and experience of a high-performing medical learner.

These findings can be considered within the context of the information-reduction hypothesis. This hypothesis submits that as people gain experience and improve their performance, they are better able to identify (and rapidly deprioritize) task-irrelevant stimuli quickly, while appropriately prioritizing task-relevant stimuli.[20,22] Further, improvement in speed and performance on a task are partially because of a reduction in the amount of information that an individual processes at the perceptual level. Newer eye-tracking research adds evidence to this theory, as domain experts have been found to have more fixations of a longer duration on task-relevant information and fewer fixations of a shorter duration on task-irrelevant information.[17]

The decreased ability to appropriately shift visual attention away from task-irrelevant stimuli may be related (at its extreme) to the concept of "helmet fire." In helmet fire, an overload of stimuli overwhelms the working memory resources of a (usually) more inexperienced physician when tasked with the management of a sick patient under high-stress conditions. Helmet fire decreases task performance and is observed in many novices in both simulation and clinical practice when managing a situation beyond their comfort level.

In our study, candidates were not specifically presented with task-irrelevant information. Task-irrelevant stimuli could be seen by some as distractors that may artificially inflate difficulty. The authors argue that, in the future, adding realistic distractors may increase the opportunity to observe discriminating behaviours that are indicative of expertise.

It may have been expected that we would have identified more positive correlations with task-relevant areas, as well as negative correlations with task-irrelevant areas. A possible explanation was found on video review. While expert avoidance of task-irrelevant stimuli is striking, consistency in viewing task-relevant data is much less obvious and systematic. Experts may take different paths to get to the same diagnostic and therapeutic destinations, and, thus, do not have chronologically consistent viewing patterns. It is also plausible that much of what makes experts proficient is their ability to effectively process information cognitively that may be challenging to discern accurately with gaze tracking.

Our study had certain limitations. As a correlational study with multiple comparisons, results should be interpreted as hypothesis generating. As a result, we limit our interpretation of this data to the general observation that higher performing residents were better able to appropriately prioritize relevant information while deprioritizing irrelevant information in simulation-based examinations. We propose that future studies design scenarios to incorporate (*a priori*) realistic task-irrelevant stimuli to more accurately simulate the clinical environment and better elucidate discriminating behaviours that might be indicative of expertise. These studies should utilize an experimental design to allow for causal conclusions about gaze tracking and performance to be made. Further, we observed poor ICC and inter-rater reliability between the two individuals

https://doi.org/10.1017/cem.2018.391 Published online by Cambridge University Press

tabulating the results and only included one external blinded assessor for performance scores for S3 and S4. Future studies that use automated computer software (as it becomes readily available) to define exactly what constitutes a visual fixation should improve the accuracy of results and speed of data generation. Finally, the authors had to make assumptions about what participants perceived based on what they looked at, without truly knowing whether their attentional resources had attended to these stimuli. Future studies could employ a mixed methods approach with post-scenario interviews to enrich the interpretation of the data.[19] Finally, our study was conducted in the simulation laboratory; therefore, we could not make conclusions on information-gathering patterns in clinical situations.

CONCLUSIONS

The study results suggest that there are certain visual behaviours in resuscitation-based simulations that are predictive of performance. These visual behaviours vary between cases because certain visual stimuli may be relevant in one patient presentation but irrelevant in another. Individuals who performed better appeared to have an improved ability to deprioritize task-irrelevant information appropriately and selectively process task-relevant information. Gaze tracking may provide educators new insights into the visual processing patterns of medical learners in simulated environments.

Acknowledgements: The authors thank Dr. Melanie Walker for her help with proofreading this manuscript.

Competing interest: The authors would like to acknowledge the Kingston Resuscitation Institute for providing funding for the research assistants and access to the eye-tracking device used in this study. None declared.

SUPPLEMENTARY MATERIAL

To view supplementary material for this article, please visit https://doi.org/10.1017/cem.2018.391

REFERENCES

1. Kim J, Neilipovitz D, Cardinal P, Chiu M, Clinch J. A pilot study using high-fidelity simulation to formally evaluate performance in the resuscitation of critically ill patients: The University of Ottawa Critical Care Medicine, High-Fidelity Simulation, and Crisis Resource Management I Study. *Crit Care Med* 2006;34(8):2167-74.
2. Watkins SC, Roberts DA, Boulet JR, McEvoy MD, Weinger MB. Evaluation of a Simpler Tool to Assess Non-technical Skills During Simulated Critical Events. *Simul Healthc* 2017;12(2):69-75.
3. Szulewski A, Howes D. Combining first-person video and gaze-tracking in medical simulation: a technical feasibility study. *The Scientific World Journal*. 2014;2014.
4. Ericsson KA, Krampe RT, Tesch-Römer C. The role of deliberate practice in the acquisition of expert performance. *Psychol Rev* 1993;100(3):363.
5. Kaufman DR, Arocha JF, Patel VL. Expertise and tacit knowledge in medicine. In: *Tacit knowledge in professional practice* (ed. Sternberg RJ, Horvath JA). Psychology Press; 1999: 89-114.
6. Szulewski A, Brindley P, Van Merrienboer JJ. Decision-making during medical crises. In: Brindley P, Cardinal P editors *Crisis Resource Management in Acute Care Medicine*, 1st ed. Ottawa: Royal College of Physicians and Surgeons Canada; 2017: 36-43.
7. Gaba DM, Howard SK, Fish KJ, Smith BE, Sowb YA. Simulation-based training in anesthesia crisis resource management (ACRM): a decade of experience. *Simul Gaming* 2001;32(2):175-93.
8. Issenberg SB, McGaghie WC, Petrusa ER, Lee Gordon D, Scalese RJ. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. *Med Teach* 2005;27(1):10-28.
9. Helle L, Nivala M, Kronqvist P, et al. Traditional microscopy instruction versus process-oriented virtual microscopy instruction: a naturalistic experiment with control group. *Diagn Pathol* 2011;6(1 Suppl 1):S8.
10. Seppänen M, Gegenfurtner A. Seeing through a teacher's eyes improves students' imaging interpretation. *Med Educ* 2012;46(11):1113-4.
11. Gegenfurtner A, Quesada-Pallarès C, Knogler M. Digital simulation-based training: A meta-analysis. *Br J Educ Technol* 2014;45(6):1097-114.
12. Hall AK, Dagnone JD, Lacroix L, Pickett W, Klinger DA. Queen's simulation assessment tool: development and validation of an assessment tool for resuscitation objective structured clinical examination stations in emergency medicine. *Simul Healthc* 2015;10(2):98-105.
13. Cook DA, Brydges R, Zendejas B, Hamstra SJ, Hatala R. Technology-enhanced simulation to assess health professionals: a systematic review of validity evidence, research methods, and reporting quality. *Acad Med* 2013; 88(6):872-83.
14. Bick JS, Demaria S Jr, Kennedy JD, et al. Comparison of expert and novice performance of a simulated transesophageal echocardiography examination. *Simul Healthc* 2013;8(5):329-34.
15. Henneman EA, Marquard JL, Fisher DL, Gawlinski A. Eye tracking: a novel approach for evaluating and improving the safety of healthcare processes in the simulated setting. *Simu Healthc* 2017;12(1):51-6.
16. Szulewski A, Roth N, Howes D. The use of task-evoked pupillary response as an objective measure of cognitive load in novices and trained physicians: a new tool for the assessment of expertise. *Acad Med* 2015;90(7):981-7.
17. Gegenfurtner A, Lehtinen E, Säljö R. Expertise differences in the comprehension of visualizations: A meta-analysis of

eye-tracking research in professional domains. *Educ Psychol Rev* 2011;23(4):523-52.

18. Szulewski A, Gegenfurtner A, Howes DW, Sivilotti ML, van Merriënboer JJ. Measuring physician cognitive load: validity evidence for a physiologic and a psychometric tool. *Adv Health Sci Educ Theory Pract* 2017;22(4):951-68.
19. Gegenfurtner A, Kok E, van Geel K, et al. The challenges of studying visual expertise in medical image diagnosis. *Med Educ* 2017 Jan51(1):97-104.
20. Haider H, Frensch PA. Eye movement during skill acquisition: more evidence for the information-reduction hypothesis. *J Exp Psychol Learn Mem Cogn* 1999;25(1):172.
21. Kok EM, de Bruin AB, Robben SG, van Merriënboer JJ. Looking in the same manner but seeing it differently: bottom-up and expertise effects in radiology. *Appl Cogn Psychol* 2012;26(6):854-62.
22. Haider H, Frensch PA. The role of information reduction in skill acquisition. *Cogn Psychol* 1996;30(3):304-37.

https://doi.org/10.1017/cem.2018.391 Published online by Cambridge University Press