To see, or not to see, that is the question:

Applying a psychological perspective to supporting medical image processing of students

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Thérése Felicitas Eder

aus Rheinfelden (Baden)

Tübingen

2021

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	06.07.2021
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter:	Prof. Dr. Katharina Scheiter
2. Berichterstatter:	Prof. Dr. Stephan Schwan

Acknowledgements

I would like to thank everyone who supported me in writing this dissertation. My thanks go to the Leibniz-Institut für Wissensmedien for providing an excellent research environment and the LEAD graduate school and research network for many opportunities to gain new insight into research and to discuss my own research.

Special thanks go to the following people who made this dissertation possible:

- Katharina Scheiter, for your supervision; for your guidance throughout this dissertation, for modeling how academic life works, for providing quick and helpful feedback to improve my texts, and for clearing organizational obstacles.
- Stephan Schwan, for being my second supervisor.
- Tamara van Gog and Hartmut Leuthold, for being on my review committee.
- Juliane Richter, for always taking time to listen to all my questions, for fruitful discussion, for looking for solutions with me when there were technical problems, and for sharing special times with me.
- Constanze Keutel and Fabian Hüttig, for broadening my view and mind for dentistry and especially for panoramic radiographs and providing insight into your expertise. Sometimes I felt like I was studying a new subject, and I am proud of the medical image interpretation skills that I learned through you.
- Nora Umbach, for racking your brain with me over complex designs and analyses which I enjoyed.
- All my colleagues for your support whenever it was needed and for sharing good and bad times.
- The dental students who participated in my studies and my student assistants for your help with data collection and coding.
- Özlem Göktürk and André Klemke, for technical support and setting up my studies.
- Leonie Jacob, Marie-Christin Krebs and Christine Postema for your feedback on my thesis.
- Simon Enz and my family for your encouragement when my thoughts were upside down and for providing me with (comfort) food.

Summary

The interpretation of medical images is an error-prone process (Pinto & Brunese, 2010) that can have serious consequences for the patients. For example, overlooked tumors can be life threatening. Also dental radiographs, which account for the largest proportion of radiographs in Germany (Bundesamt für Strahlenschutz, 2016), can contain such serious anomalies, for example, calcifications of the carotid artery (Friedlander et al., 2005). When physicians fail to identify anomalies, the errors may result from not looking at the anomalies (detection error), not recognizing the features of an anomaly (recognition error), or deciding against the relevance of suspicious features of the anomaly (decision error) (Kundel et al., 1978; Wu & Wolfe, 2019).

To date, there are only few evaluated training methods to improve medical image processing and resulting diagnostic performance, and the evaluation of specific training methods is lacking (Kok et al., 2017). Therefore, this dissertation aims at developing and evaluating different training methods to support dental students in reading panoramic radiographs (Orthopantomogramm; OPT). Three studies evaluated three training methods that were expected to result in the improved detection of anomalies and more intense visual processing. This intense visual processing should be reflected in sooner, longer and more frequent fixations on anomalies. Unexpectedly, only the training method in Study 2 improved the detection of anomalies and none of the training methods led to the expected intensification of visual processing.

Study 1 examined an individualized full coverage training to help dental students to search in all areas of the OPTs and thereby reduce the number of missed anomalies. Dental students either received the training for five OPTs in the intervention group (n = 38) or diagnosed five OPTs in the control group (n = 23) in a pre- and post-test setting. The training consisted of gaze feedback comparisons with eye movement visualizations of a peer model showing full coverage and their own gaze behavior. The results showed only small and not meaningful improvements in the detection rate for anomalies. Similarly, the training had a very small positive effect on the visual coverage rate. Gaze behavior regarding anomalies changed with training towards expanded visual search with shorter and fewer fixations on anomalies. The time to first fixation indicated a minor shift in attention towards anomalies located in the periphery. An exploratory analysis revealed that the dental students made five times more recognition and decision errors compared to detection errors, suggesting that detection errors addressed in this training are only a small part of the problem.

Study 2 evaluated training that aimed to reduce recognition and decision errors by focusing on anomaly identification. In a crossover design, two training sessions that addressed either anomalies located in the periphery, for example maxillary sinuses and the neck, (peripheral anomalies) or the dentition (central anomalies) were tested simultaneously. In one group, dental students ($n = 39$) first received training for the recognition of peripheral anomalies and second training for the recognition of central anomalies, while the other group ($n = 39$) received the training sessions in the reversed order. The training method compared two OPTs with and without disease and two OPTs with the same disease. Additionally, colored highlights tagged the relevant areas in the OPTs and the instruction contained a verbal description of the characteristic features of anomalies. The results showed that the training was effective by improving diagnostic performance. The order of the training sessions seemed to affect the effectiveness and the learning times of the training seemed to influence the output. The training did not change the visual search behavior.

Study 3 investigated training with eye movement modeling examples (EMME) designed to combine visual search strategies and object identification. In the intervention group, dental students ($n = 42$) saw three EMME videos from experts with didactical verbal explanations between the pre- and the post-tests. Dental students in the control group ($n = 41$) only performed the pre- and the post-test. In an online study of 31 dental students, the study was replicated with a second evaluation of the training without measuring eye movements. However, the training did not improve the detection of anomalies in either of the experiments. Students' visual search behavior did not change for visual coverage rate, but the intervention led to shorter, fewer, and later fixations on anomalies. Exploratory analysis confirmed the findings from Study 1 on the distribution of error types with less than 20% detection errors.

The results of these studies suggest that the training method which focuses on anomaly identification (Study 2) is effective when recognition and decision errors dominate. This training method provides knowledge about the visual properties of anomalies and their discrimination (Study 2). In contrast, training methods that teach visual search strategies (Studies 1 and 3) do not appear to be beneficial under these circumstances. However, further research is needed to investigate possible long-term effects of search strategy training (Kramer et al., 2019). Changes in eye movements indicate that the training may trigger a change in cognition that could lead to improved diagnostic performance after some time. This dissertation implies that the type of errors needs to be considered before applying training in dental education. Comparison of radiographic images, as examined in Study 2, provides a supportive

training method that could be easily implemented in university teaching and improves diagnostic performance.

Zusammenfassung

Bei der Interpretation medizinischer Bilder handelt es sich um einen fehleranfälligen Prozess (Pinto & Brunese, 2010), der schwerwiegende Folgen für die Patienten haben kann, wenn zum Beispiel Tumore übersehen werden. Auch zahnärztliche Röntgenaufnahmen, die in Deutschland den größten Anteil ausmachen (Bundesamt für Strahlenschutz, 2016), können schwerwiegende Anomalien beinhalten, wie z. B. Verkalkungen der Halsschlagader (Friedlander et al., 2005). Wenn Ärzte Anomalien nicht identifizieren, können die entsprechenden Fehler dadurch zustande kommen, dass sie die Anomalie nicht betrachten (Entdeckungsfehler), die Merkmale einer Anomalie nicht erkennen (Erkennungsfehler) oder sich gegen die Relevanz verdächtiger Merkmale entscheiden (Entscheidungsfehler) (Kundel et al., 1978; Wu & Wolfe, 2019).

Bislang gibt es nur wenige evaluierte Trainings zur Verbesserung der medizinischen Bildverarbeitung und der daraus resultierenden diagnostischen Leistung und es fehlt die Evaluation spezifischer Trainingsmethoden (Kok et al., 2017). Ziel dieser Dissertation war es daher, verschiedene Trainingsmethoden zur Unterstützung von Zahnmedizinstudierenden beim Lesen von Panoramaröntgenaufnahmen (Orthopantomogramm; OPT) zu entwickeln und zu evaluieren. In drei Studien wurden drei Trainingsmethoden evaluiert, die zu einer verbesserten Erkennung von Anomalien und zu einer intensiveren visuellen Verarbeitung führen sollten. Diese intensivere Verarbeitung sollte sich in Blickbewegungen mit längeren, häufigeren und früheren Fixationen auf Anomalien widerspiegeln.

Studie 1 untersuchte ein individualisiertes Training zur vollständigen Abdeckung bei der visuellen Suche. Dieses Training sollte Zahnmedizinstudierende dazu ermutigen, in allen Bereichen der OPTs zu suchen und dadurch die Anzahl der übersehenen Anomalien zu reduzieren. Insgesamt erhielten 61 Zahnmedizinstudierende entweder das Training für fünf OPTs in der Interventionsgruppe oder diagnostizierten fünf OPTs in der Kontrollgruppe in einem Prä- und Posttest-Design. Das Training bestand aus Vergleichen von Blickbewegungsvisualisierungen eines Peer-Modells mit einer vollständigen Abdeckung bei der visuellen Suche und dem eigenen Blickverhalten. Die Ergebnisse zeigten nur geringe und nicht bedeutsame Verbesserungen der Erkennungsrate von Anomalien. Ebenso wirkten sich die Trainings in sehr geringem Maße positiv auf die visuelle Abdeckung aus. Das Blickverhalten bezüglich der Anomalien veränderte sich durch das Training hin zu einer erweiterten visuellen suche mit kürzeren und weniger Fixationen auf Anomalien. Die Zeit bis zur ersten Fixation deutet auf eine kleine Verschiebung der Aufmerksamkeit in Richtung Anomalien, die in der

Peripherie lokalisiert waren, hin. Eine explorative Analyse ergab, dass Zahnmedizinstudierende fünfmal mehr Erkennungs- und Entscheidungsfehler im Vergleich zu Entdeckungsfehlern machen, was darauf hindeutet, dass die in diesem Training angesprochenen Entdeckungsfehler nur einen kleinen Teil des Problems darstellen.

Studie 2 evaluierte ein Training, das Erkennungs- und Entscheidungsfehler adressierte, indem es bei der Identifikation von Anomalien ansetzte. In einem Crossover-Design wurden zwei Trainings gleichzeitig getestet. Sie behandelten entweder Anomalien die sich in der Peripherie (periphere Anomalien), wie beispielsweise den Kieferhöhlen oder dem Hals, oder in der Mundhöhle (zentrale Anomalien) befinden. In einer Gruppe erhielten die Zahnmedizinstudierenden (n = 39) zuerst das Training zur Entdeckung peripherer Anomalien und anschließend das Training zur Entdeckung zentrale Anomalien, während die andere Gruppe (n = 39) die Trainings in umgekehrter Reihenfolge erhielt. Als Trainingsmethode wurden Vergleiche von zwei OPTs mit und ohne Erkrankung und Vergleiche von zwei OPTs mit der gleichen Erkrankung verwendet. Zusätzlich wurden die relevanten Bereiche in den OPTs farblich hervorgehoben und eine verbale Beschreibung der charakteristischen Merkmale der Anomalien präsentiert. Die Ergebnisse zeigten, dass das Training effektiv war, indem es die diagnostische Leistung verbesserte. Dabei scheint die Reihenfolge des Trainings einen Einfluss auf die Effektivität zu haben. Außerdem liegt ein Zusammenhang der Effektivität mit den Lernzeiten des Trainings nahe. Das Training veränderte das visuelle Suchverhalten nicht.

In Studie 3 wurde ein Training mit Eye Movement Modeling Examples (EMME) untersucht, das visuelle Suchstrategien und Objektidentifikation kombinieren sollte. In der Interventionsgruppe sahen die Zahnmedizinstudierenden (n = 42) zwischen dem Prä- und Posttest drei EMME-Videos von Experten mit didaktischen verbalen Erklärungen. Die Zahnmedizinstudierenden in der Kontrollgruppe (n = 41) führten nur den Prä- und Posttest durch. In einer Online-Studie mit 31 Zahnmedizinstudierenden replizierte ich die Studie mit einer zweiten Evaluation des Trainings ohne Messung der Augenbewegungen. Allerdings verbesserte das Training in beiden Experimenten nicht die Erkennung von Anomalien. Das visuelle Suchverhalten der Studierenden änderte sich nicht für die visuelle Abdeckungsrate, aber die Intervention führte zu kürzeren, weniger und späteren Fixationen auf Anomalien. Eine explorative Analyse bestätigte die Ergebnisse aus Studie 1 zur Verteilung der Fehlertypen mit weniger als 20% Erkennungsfehlern.

Die Ergebnisse dieser Studien deuten darauf hin, dass Trainingsmethoden, die die Identifikation von Anomalien adressieren (Studie 2), effektiv sind, wenn Erkennungs- und Entscheidungsfehler dominieren. Dabei vermitteln diese Trainings Wissen über die visuellen

Eigenschaften der Anomalien und ihrer Unterscheidung (Studie 2). Hingegen scheinen Trainingsmethoden, die lediglich Sehstrategien für die visuelle Suche vermitteln (Studie 1 und 3), unter diesen Umständen nicht förderlich zu sein. Weitere Forschung ist jedoch notwendig, um mögliche Langzeiteffekte von Suchstrategietrainings zu untersuchen (Kramer et al., 2019). Veränderungen in den Augenbewegungen deuten darauf hin, dass die Trainings möglicherweise eine Veränderung der Kognition auslösen, die nach einiger Zeit zu einer verbesserten diagnostischen Leistung führen könnte. Diese Arbeit impliziert für die Praxis, dass vor der Anwendung von Trainings die Art der Fehler berücksichtigt werden sollte. Der Vergleich von Röntgenbildern, wie er in Studie 2 untersucht wurde, bietet eine unterstützende Trainingsmethode, die leicht in die universitäre Lehre umsetzbar wäre und die diagnostische Leistung verbessert.

List of Publications and Contributions

The manuscripts of Study 1 and 2 have been published elsewhere. This list contains all publications of this thesis and the proportional contribution of all authors is listed below.
The publications of Study 1 and 2 and the submitted Manuscript of Study 3 are presented in chapters 7-9.

Manuscript of Study 1

Author	Author position	Scientific ideas %	Data generation %	Analysis & Interpretation %	Paper writing %
Thérése Eder	First author	40%	80%	90%	60%
Juliane Richter	Second author	5%	15%	5%	10%
Katharina Scheiter	Third author	40%	0%	5%	10%
Constanze Keutel	Forth author	5%	0%	0%	5%
Nora Castner	Fived author	0%	5%	0%	5%
Enkelejda Kasneci	Sixth author	5%	0%	0%	5%
Fabian Huettig	Seventh author	5%	0%	0%	5%
Title of paper:		How to support dental students in reading radiographs: effects of a gaze-based compare-and-contrast intervention			
Status in publication process:		Published: Eder, T. F., Richter, J., Scheiter, K., Keutel, C., Castner, N., Kasneci, E., & Huettig, F. (2021). How to support dental students in reading radiographs: effects of a gaze-based compare-and-contrast intervention. *Advances in Health Sciences Education,* 26(1), 159-181. https://dx.doi.org/10.1007/s10459-020-09975-w			

Manuscript of Study 2

Author	Author position	Scientific ideas %	Data generation %	Analysis & Interpretation %	Paper writing %
Thérése Eder	First author	70%	95%	90%	70%
Juliane Richter	Second author	5%	5%	5%	5%
Katharina Scheiter	Third author	15%	0%	5%	15%
Fabian Huettig	Forth author	5%	0%	0%	5%
Constanze Keutel	Fifth author	5%	0%	0%	5%

Title of paper:	Comparing radiographs with signaling improves anomaly detection of dental students: An eye-tracking study.
Status in publication process:	Published: Eder, T. F., Richter, J., Scheiter, K., Huettig, F., & Keutel, C. (in press). Comparing radiographs with signaling improves anomaly detection of dental students: An eye-tracking study. *Applied Cognitive Psychology.* https://doi.org/10.1002/acp.3819

Manuscript of Study 3

Author	Author position	Scientific ideas %	Data generation %	Analysis & Interpretation %	Paper writing %
Thérése Eder	First author	45%	100%	95%	70%
Katharina Scheiter	Second author	50%	0%	5%	15%
Juliane Richter	Third author	5%	0%	0%	5%
Constanze Keutel	Forth author	0%	0%	0%	5%
Fabian Huettig	Fifth author	0%	0%	0%	5%

Title of paper:	I see something you don't: Eye movement modeling examples do not improve anomaly detection in interpreting medical images.
Status in publication process:	Revision

Table of contents

1. Introduction

Interpreting medical images such as radiographs is an error-prone process even in experts (Pinto & Brunese, 2010). For the patient, misdiagnosis can cause pain or even threaten life, for example when radiologists overlook cancerous tumors. Therefore, every effort should be made to prevent or at least minimize misdiagnosis. Whereas in most medical disciplines specialized radiologists perform the interpretation of radiographs, dentists take on this task themselves. Reading radiographs is part of their daily work. In Germany, dental radiographs account for the largest share of radiographs taken (43%) (Bundesamt für Strahlenschutz, 2016). Dentists can also detect serious findings in panoramic radiographs of the jaw (orthopantomograms; OPTs) which show not only the oral cavity with the teeth but also the maxillary sinuses up to the neck. Dentists are required to also inspect the peripheral areas around the oral cavity and detect possible anomalies, such as calcifications of the carotid artery, which can potentially lead to a stroke (Friedlander et al., 2005; Tamura et al., 2005). Thus, high error rates of 41% can weigh heavily in the interpretation of dental radiographs (Stheeman et al., 1996). To reduce such errors, it is important to understand the processes involved in the visual search for anomalies in radiographs. Thereby, bottom-up processes resulting from the perception of the stimulus (e.g., radiographs) and top-down processes resulting from knowledge-based decision making play a crucial role.

Psychological methods and a psychological understanding of the perceptual and cognitive processes involved in medical image processing and decision making for diagnosis are a foundation for research on this topic. The findings from such studies investigating the perception and processing of visual search in medical images form a basis for the development of training to support the processes (Jensen et al., 2008). Most previous studies have looked at expertise differences in visual search on radiographs to extract characteristics of a successful search (e.g., Donovan & Litchfield, 2013; Kundel et al., 2007; Nodine et al., 1999).

It has proven useful to apply eye tracking as the method of investigating visual perceptual processes. With this method, it was possible to find out, for example, that novices perform an insufficient visual search compared to experts and intermediates because they process smaller parts of the images and look less often at the relevant areas (Jaarsma et al., 2014). Accordingly, eye tracking is an optimal tool for evaluating search patterns and strategies and can be used to visualize them. Furthermore, eye tracking is not only used for evaluating visual search, but also offers the possibility to use eye movement visualizations as an intervention tool for training (gaze-based intervention) (Kok & Jarodzka, 2017a).

To improve patient care, it is important to hone the relevant skills in university teaching and improve existing approaches with evaluated high-quality training methods. However, evaluation of these courses, evaluated training methods, and research on the development and acquisition of medical image processing skills necessary to diagnose radiographs are lacking (Gegenfurtner et al., 2011; Kok et al., 2017). Little is also known in dentistry about the underlying processes of interpreting radiographs such as the perception and detection of anomalies. To date, there is a lack of evaluated training methods for use in university teaching that address visual search to improve the interpretation of radiographs (Kok et al., 2017), particularly dental radiographs.

There is an urgent need for evaluated training methods to support dental students in reading radiographs at an early stage of their career. Therefore, the following questions arise: How can dental students be supported in evaluating panoramic radiographs? What processes do training methods need to address in order to achieve improved diagnostic performance? This dissertation addresses these questions by studying how gaze-based and instructional interventions can aid dental students in their interpretation of radiographs. The analysis of these research questions provides deep insight into how effective training methods should be developed for university teaching. Thus, this work will hopefully contribute to more accurate diagnosis and better treatment of patients.

In what follows, an overview on clinical reasoning is provided before presenting eye tracking as a tool to evaluate medical image processing. After discussing the visual search process and errors in medical imaging, an overview of training methods for medical image processing follows. The three studies are presented in chapters 7, 8, and 9. The results of the subsequent studies evaluating three training methods are then discussed and the strengths, limitations, and implications of this work are presented.

2. The process of clinical reasoning and its development

Imagine going to the dentist because you have toothache. The dentist will look at your medical history and ask you, for example, which tooth hurts and for how long it has been painful. Then, the dentist conducts some examinations and finally arrives at a diagnosis and presents you with treatment options, one of which is later implemented. During this whole process, the dentist performs clinical reasoning.

The concept of clinical reasoning originates from medicine and has a wide range of definitions: It is seen as being synonymous with "problem solving, decision making, [or] judgment" by physicians or medical students in making a diagnosis (Norman, 2005, p. 418). Clinical reasoning is also described as behavior as well as process, ability, or outcome in the context of coming to a clinical diagnosis (Schuwirth et al., 2020). Thus, clinical reasoning encompasses all the cognitive and behavioral processes that lead to a clinical diagnosis and further treatments and the diagnosis itself (cf. Schuwirth et al., 2020). In summary, clinical reasoning refers to the thinking and judgment processes involved in making diagnostic and therapeutic decisions based on information about the patient, which requires knowledge of the disease and skills of the physician. In such clinical reasoning processes, physicians “collect cues, process the information, come to an understanding of a patient problem or situation, plan and implement interventions, evaluate outcomes, and reflect on and learn from the process” (Levett-Jones et al., 2010, p.516). This logical process does not reflect a linear reasoning, but a cyclical one, so that reflection on the process influences the cue collection next time (Levett-Jones et al., 2010).

These processes of clinical reasoning, which originate from medicine, are highly interesting to psychological research. This is a knowledge-rich domain that enables fundamental research questions to be addressed in highly relevant, applied field. Therefore, there is also a lot of psychological research that has dealt, first, especially with the cognitive processes and, second, with the perceptual processes involved in clinical reasoning (Gruppen, 2017). Cognitive processes play a role, for example, when processing the information from the examinations and deciding on a diagnosis or treatment. Perceptual processes are necessary when collecting cues for a disease from examinations of the patient, such as observations from radiographs, computer tomography or electrocardiograms.

On the whole, clinical reasoning, despite its roots in the medical field, has been of interest to psychology for a long time (Schmidt & Mamede, 2020). Reasons are the knowledge-rich problem-solving process, the application of psychological aspects on medical education and the

prototypical case of medical image processing for visual processes and high-level performance. In the following three sections, these aspects relevant to psychological research are discussed.

2.1 Knowledge-rich processes of clinical reasoning

First, regarding clinical reasoning as a knowledge-rich process, psychological research can help to predict and investigate the underlying cognitive aspects that are necessary to successfully apply knowledge in clinical reasoning.

One result of the interdisciplinary research of psychology in this medical field is the theory of knowledge encapsulation, which applies the concepts of scripts and schemas to clinical reasoning (Norman, 2005). The theory of knowledge encapsulation describes how clinical reasoning develops in medical professionals in three stages (Boshuizen & Schmidt, 1992, 2008). Thereby, the gradual development of schemata and knowledge organization is shown. At the first stage, medical students establish a causal network to acquire medical knowledge. They learn a lot of basic science - biomedical knowledge - which needs to be structured and stored in a network for successful reasoning. The networks contain nodes representing concepts and connections between different higher or lower-order concepts. Higher-order concepts (general description of disease) contain specific lower-order concepts as symptoms and signs (cf. Jarodzka et al., 2013). At the end of the first stage, medical students repeatedly use connections between concepts so that they become automated. Thereby, intermediate concepts between start and end concepts are skipped. This process is called knowledge encapsulation and is the main component of the second stage. In the second stage, biomedical knowledge is encapsulated/integrated into clinical knowledge (Boshuizen & Schmidt, 2008). Medical students incorporate their experience from patient contacts and diagnosis into the network. Thus, basic knowledge is converted into simple causal models for signs and symptoms. Encapsulation increases efficiency and ensures that only knowledge relevant to the particular case is activated in the networks. In the transition to the third stage, the networks are transformed into a different form of knowledge organization, that is, illness scripts. Illness scripts contain three components: First, the conditions for a disease such as personal, environmental, or hereditary factors. Second, the pathophysiological processes of the disease (also encapsulated) and third, the signs and symptoms of the disease. The advantage of the illness script over the networks is that the entire script is always activated, and the individual elements of the script can be retrieved automatically. Thereby, medical professionals can reason very efficiently. Evidence for the theory of encapsulation of knowledge has been found in many studies (for an overview see Boshuizen et al., 2020).

It takes a long time to pass these steps and become a medical expert, who is able to combine and masters various relevant skills and knowledge (Ericsson, 2004; Norman et al., 2006). It is essential to practice with many cases to achieve a high level of clinical reasoning skills (Ericsson et al., 1993). For example, mammographers' interpretive skills increase with the number of cases they have processed (Nodine & Mello-Thoms, 2000). Experienced mammographers read between 9,459 and 12,145 cases. This high number of cases that must be mastered for high performance is also known from other fields (e.g., chess, performing music, cf. Ericsson et al., 1993). Thus, frequent exposure to cases and practice is necessary to show high domain-specific performance in a task.

2.2 Medical education

Second, psychological methods and findings can be used to examine, evaluate, and support medical education and the development of clinical reasoning skills.

Before physicians are exposed to many cases in their professional lives, they undergo initial training during their medical studies at university. Medical students learn basic biomedical knowledge (e.g., biochemical knowledge of pathophysiology) and study theoretical aspects of diseases and treatments (cf. Norman, 2005). University education aims to ensure that medical students use their knowledge to make correct decisions about diagnoses and treatments, and so clinical reasoning begins to develop during medical education. Although the university curriculum includes internships in a variety of medical settings, most hands-on learning experiences follow graduation. So, this question remains: How should medical content be taught during medical education to support students' initial knowledge acquisition that will enable them to build high levels of professional knowledge and performance in their careers? Traditionally, two directions of problem-based learning dispute about this question in medical education (Servant-Miklos, 2019). One direction focuses on teaching processes, referring to the theoretical view of clinical reasoning as a problem-solving process. The other direction focuses on teaching knowledge, addressing the theoretical aspects of knowledge organization for clinical reasoning. First studies showed that the teaching knowledge approach appears to be more effective in general (Monteiro et al., 2020; Schmidt & Mamede, 2015). However, it is not clear what medical teaching should look like to achieve a high-quality performance (Schmidt & Mamede, 2015).

Cognitive psychology can contribute to transform and support medical education (Schmidt & Mamede, 2020). Schmidt and Mamede (2020) describe interventions based on cognitive theories that can make the development of medical professional skills more effective

(fostering self-explanation, reducing cognitive load, supporting distributed, retrieval, and interleaving practice) and state that the interventions fit with methods already used in medical education, such as worked examples, team-based or problem-based learning. Nevertheless, further research is needed to study and apply the cognitive interventions in practice. This dissertation addresses such interventions and explores how the initial learning process in medical education can be supported so that knowledge is transferred and applied through targeted learning opportunities for medical image processing (introduced in the next section).

2.3 Processing of medical images

The third reason why clinical reasoning has been of interest to psychologists is that medical image processing provides a real-world example of how visual processing contributes to high-level performance. Clinical reasoning in many situations requires that a physician in order to arrive at a diagnostic decision will process (i.e., inspects and interprets) medical images (e.g., radiographs, electrocardiograms or microscopic images). From a psychological perspective, medical image processing heavily relies on bottom-up (i.e., stimulus-driven) and top-down (i.e., knowledge-driven) attentional processes, which interact and may also be the source of diagnostic errors (Ganesan et al., 2018; Jarodzka, Boshuizen, et al., 2013). Clinical reasoning and medical image processing in particular have been studied in the context of psychological research with the aim to find out more about the cognitive characteristics of high-performing physicians and about the development of skills that are relevant for the visual search in medical images. Before going into the details of the psychological perspective of medical image processing (chapter 4), the next chapter gives an excursus on the most commonly used method to study visual search in medical images, that is eye tracking.

3. Eye tracking: A tool to study medical image processing

Eye tracking represents a central method to study medical image processing used by the majority of studies investigating the visual search in medical images (e.g. Carmody et al., 1984; Donovan & Litchfield, 2013; Grünheid et al., 2013; Jaarsma et al., 2014, 2015; Kok et al., 2012, 2016; Kundel et al., 2007; Kundel & La Follette, 1972; Manning et al., 2006; Turgeon & Lam, 2016). The following provides a brief introduction to eye tracking before giving an overview on the research of medical image processing that used this method in the next chapter.

Eye tracking is a technology that enables us to measure eye movements of a person or even animals (e.g. Somppi et al., 2012) with a camera. During visual perception, light enters the eye through the pupil and humans see information that falls on the fovea in focus (foveal vision) (Holmqvist et al., 2011). Humans can also recognize movements or high contrast information that fall in the periphery of the retina, which is called peripheral vision (cf. Kok & Jarodzka, 2017a). Eye tracking captures only foveal vision and thus provides information only about the areas that were looked at directly. Therefore, the pupil and corneal reflection method is used to measure the eye movements (Holmqvist et al., 2011). A camera detects the position of the pupil, as the darkest point of the eye, and the position of a corneal reflection caused by an infrared light, as the brightest point of the eye. When the eye moves, the distance between the pupil and the corneal reflection changes, making it possible to measure eye movement. In contrast, eye tracking cannot measure when an observer perceives an object, such as an anomaly, by peripheral vision without looking at it directly (Kok & Jarodzka, 2017a). This also imposes limitations on the interpretation of eye movement data.

Eye movements are classified into fixations, which describe the gaze focusing on a specific area without moving, and saccades, which describe the movements between fixations (Holmqvist et al., 2011). During saccades the observer does not absorb information from the stimulus. Instead, information can be processed only during fixations, when the eyes are positioned on a specific area. Based on eye movement data, inferences about a person's cognitive processes can be derived (Just & Carpenter, 1980). Just and Carpenter (1980) stated basic assumptions about the relationship between eye movements and cognitive processes. They assume that visual stimuli are the focus of interest as long as they are being looked at (eye-mind assumption). Perceived visual stimuli are processed immediately at the cognitive level (immediacy assumption). Consequently, visual information of the stimuli is processed cognitively if and as long as a person fixates the stimulus.

Even though there are limitations to this straightforward interpretation of eye movements reflecting cognitive processes (e.g., Smith et al., 2017), “looking at information is a necessary condition for being processed cognitively; it is, however, not sufficient” (Kok & Jarodzka, 2017b, p. 1). This means that when observers perceive visual information, they must necessarily look at it, although this does not predict how they will cognitively process the information. Eye movement measures can be opposite in expression (e.g. long vs. short fixation durations) and still both represent higher processing skills (c.f. Jarodzka & Boshuizen, 2017). Reasons for this ambiguity are the specificity of task and stimuli. Such ambiguities have been found for other processing measures as well. For example, time to complete a task should be positively related to performance (Wickelgren, 1977), but negative relationships have been found with reading performance as ability level increases (Goldhammer et al., 2014). Thus, the inferences from eye movements to cognitive processes should be drawn cautiously and preferably with a theoretical basis (Kok, 2019)

In this paragraph, commonly used eye-tracking measures for visual search in medical images are presented. The measures often refer to a defined area of interest (AOI) in the medical images, which in this context usually includes an anomaly. Van der Gijp et al. (2017) identified the following relevant measures: Time to first fixation with respect to an AOI, total fixation duration/time (on AOIs), number of fixations (on AOIs), number and length of saccades, and image coverage (van der Gijp et al., 2017). The time until a person looks at an AOI (anomaly) for the first time – time to first fixation – reflects the person’s ability to quickly detect the anomaly. Higher numbers of fixations as well as longer fixation times on AOIs indicate more intensive processing of the AOI, e.g. due to its relevance to the task. Long saccades together with high frequencies of fixations typically indicate intense visual search behavior. Finally, image coverage refers to the degree to which a person inspected an image by fixating in multiple areas.

In general, eye tracking can serve as a tool to investigate conditions for processing information and give insight into visual search processes as in medical image processing (Kok & Jarodzka, 2017b).

4. Processing of medical images: The psychological perspective

While the last chapter introduced the method eye tracking for investigating visual search in medical images, this chapter addresses the theoretical concepts for visual search, empirical studies that use eye tracking to investigate visual search and the problems that may arise when processing medical images. In the following, the processing of medical images is understood as the visual search for signs of diseases in medical images, such as radiographs, microscopic images or electrocardiograms, and the related cognitive processes (e.g. for diagnostic decision-making). This entire processing of medical images with its perceptual and cognitive processes is a very interesting one from a psychological point of view. Psychological methods can be used to examine the individual processing steps, their interaction and any problems that may arise from perceptual and cognitive processes.

The second part of this chapter deals with the problems that arise and how errors occur in the interpretation of medical images. The first part explores the question of how physicians process medical images with theoretical concepts and empirical studies. First, the global-focal model (Nodine & Mello-Thoms, 2000) which focuses on the perceptual process of visual search in medical images is discussed. Second, visual search and its characteristics in dental radiographs is described. Third, the model of cognitive skills operating in medical diagnosis that provides a basis for models of medical image processing is presented (Jarodzka, Boshuizen, et al., 2013; Jarodzka & Boshuizen, 2017). Fourth, an introduction follows of a framework of the knowledge and skills that are necessary to evaluate medical images and to process medical images successfully (van der Gijp et al., 2014).

4.1 Visual search in medical image processing

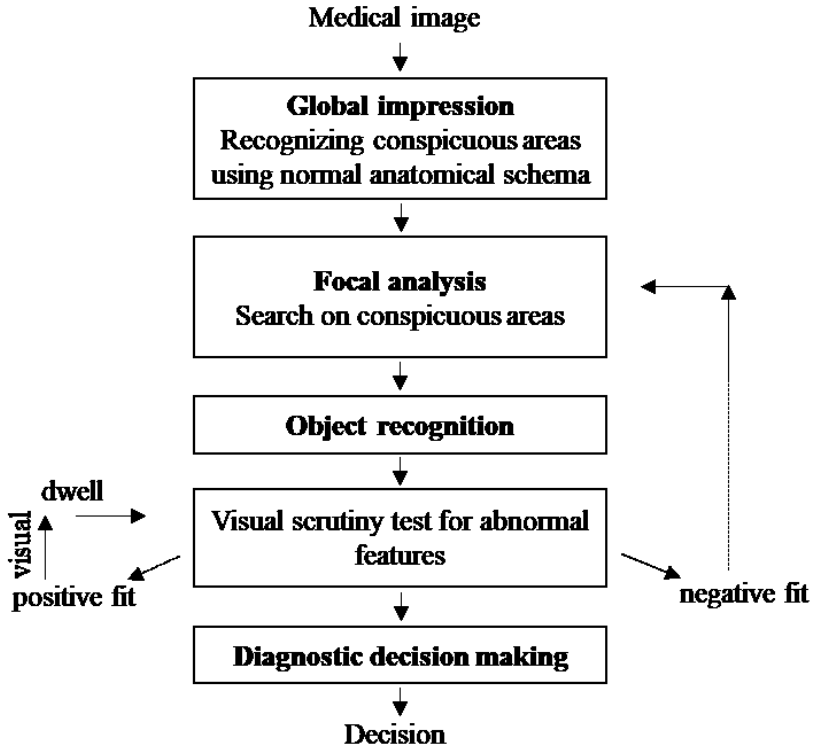
4.1.1 Global and focal processing of medical images

The visual search process in medical images is described in the global-focal models (Kundel et al., 2007; Nodine & Kundel, 1987; Nodine & Mello-Thoms, 2000). The models focus on two elements: the global impression and the focal search. In the following, the general Perceptual Model of the Radiology Task (Nodine & Mello-Thoms, 2000) is presented in more detail (see Figure 1). The model captures the process of visual search in medical images, starting with the observation of medical images and ending with a diagnostic decision. It assumes that an observer receives a global impression when first looking at the image. In the global impression,

the observer uses an anatomical scheme (i.e., a representation about the arrangement of components in the human body) to identify regions that are free of anomalies or could represent anomalies. The regions identified as containing possible anomalies are examined in more detail in the focal analysis: The observer examines the detected object to extract features relevant to the disease. If such features are present, the observer looks at the objects longer and comes to a diagnostic decision. In the diagnostic decision process, the criteria of the possible diagnosis are matched with the features found. If no features relevant to the disease are found, the search process starts again at the focal analysis stage for a new abnormal region. In this model, bottom-up processes are involved in observing medical images and top-down processes are involved in diagnostic decision making.

Figure 1

Nodine's and Mello-Thoms's Perceptual Model of the Radiology Task (Nodine & Mello-Thoms, 2000, p. 869).



Note. Reprinted from the Handbook of Medical Imaging, Nodine & Mello-Thoms, *The nature of expertise in radiology,* p. 869, Copyright (2000), with permission from SPIE and the author Claudia Mello-Thoms.

While experts can use the global impression to immediately detect anomalies, novices cannot use the first global impression and must instead use a search-to-find method from the

beginning (Kundel et al., 2007). Thus, the application of the global impression and the focal search differs according to the level of visual expertise (Nodine & Mello-Thoms, 2000), which is defined as "reproducibly superior visual skill when making a diagnosis from a medical image" (Gegenfurtner et al., 2017, p. 98).

Evidence for the existence of the first global impression comes from studies using eye movements to evaluate how fast experts look at anomalies. Kundel et al. (2007) found that observers detect anomalies in mammograms more frequently when they looked at them sooner and showed fast global processing. This relationship between speed and performance with increasing expertise level was also found in the study of Nodine et al. (1999) investigating observers at three expertise levels in mammography. Additionally, in the field of chest radiographs, evidence for the global processes were found in wider views and less time on tasks for the more experienced observer (Manning et al., 2006). In contrast, in another study, the experts did not look at the anomalies in the chest radiographs earlier than other observers, and so the study found no evidence for global processing (Donovan & Litchfield, 2013). However, the sample size of these studies was relatively small because experts are rarely found.

A meta-analysis of Gegenfurtner et al. (2011) summarized visual search with eye-tracking studies in different professional domains (e.g. sports, transportation, and medicine) and provided evidence for the global-focal processes. Experts in comparison to non-experts fixated on relevant areas sooner and more often, fixated less on redundant areas, showed shorter fixation durations in general, longer saccades and better parafoveal processing. These results indicate the global processing which could be only used by experts. However, the eye tracking measures in visual search highly depended on the task, characteristics of the image and the domain (Gegenfurtner et al., 2011).

In the domain of medical images, a scoping review by Al-Moteri et al. (2017) investigated visual cue processing in medical decision making of eye-tracking studies and also found hints for global-focal processes. Further evidence especially for radiological images is provided by a systematic review by van der Gijp et al. (2017). Van der Gijp et al. (2017) investigated differences between experts and novices regarding their eye movements during medical image processing. They found that experts typically show sooner fixations on AOIs, make less fixations in general and longer saccades which reflect the efficient global processing. The results regarding visual coverage were inconclusive. One reason for this could be that only few studies investigated visual coverage (four studies). Furthermore, van der Gijp et al. (2017) found differences at expertise level in fixation duration on AOIs and the number of fixations on AOIs depending on the task. An example: Whereas observers fixate shorter on AOIs with

increasing expertise levels in recognition-only tasks, they fixate longer when the task combines recognition and interpretation (cf. van der Gijp et al., 2017). The task dependency is also in line with the results of the meta-analysis mentioned above (Gegenfurtner et al., 2011). Due to the dependency of the visual search on the specific task and domain, the question arises whether the previous findings can be applied to dental radiographs.

4.1.2 Visual search in dental radiographs

So far, little is known about eye movements in dental medicine images and the intermediate level of dental students. When considering the different characteristics of OPTs compared to, for example, chest radiographs, it seems implausible that the exact same processes are taking place. For instance, chest radiographs typically show only a small number of up to five anomalies (Donovan & Litchfield, 2013; Kundel et al., 1978). In contrast, the OPTs used in these dissertation studies contained up to 26 anomalies. Diagnosing OPTs relies on a hybrid search for multiple different targets, thereby requiring observers to know all the characteristics of potential targets and match those to the actual visual characteristics of radiographs (Wolfe, 2012). The occurrence of multiple targets is known to complicate visual search and makes it less effective (Wolfe et al., 2016). Moreover, in OPTs anomalies are also found in the peripheral areas of the jawbone or in the maxillary sinus or are part of the radiographs as superimpositions of soft and hard tissue in the vicinity of the oral cavity. These complex anatomical areas with maxillary sinuses, spine with neck, temporomandibular joints and the jaws with dentition overlapping in the 2D format makes the interpretation challenging (Bahaziq et al., 2019). The peripheral areas may include various secondary findings of general medical relevance (oncology, cardiovascular disease) that require referral to other specialists for further diagnosis. Anomalies located in the periphery are often of low prevalence (Constantine et al., 2018; Vallo et al., 2010), and the level of prevalence of anomalies affects visual search (Wolfe, 2016). A first study with non-experts showed that lower prevalence led to more missed targets in an artificial baggage screening task (Wolfe et al., 2005). Thus, it might be more difficult to detect anomalies with low prevalence such as those in the periphery of OPTs.

The following part summarizes the previous results of visual search in dental radiographs. Hermanson et al. (2018) found in a pilot study that dental observers mostly first fixate on areas of high contrast/salience as radiopaque or radiolucent. Additionally, the observers used tooth-by-tooth scanning in the periapical radiographs, which normally display two to three teeth only.

In another study, more and less experienced dentists (more/less than five years of experience) assessed panoramic radiographs (Grünheid et al., 2013). In general, more

experienced dentists looked for a shorter period of time at the radiograph and showed a systematic scanning pattern whereas less experienced dentists did not show a systematic scanning pattern but a higher coverage in scanning the radiograph and fixated anomalies more often than more experienced dentists. These results are in line with a global processing for more experienced dentists with less fixations, shorter viewing time and a systematic scanning pattern.

Turgeon and Lam (2016), who also investigated panoramic radiographs in dental students and radiologists (oral and maxillofacial radiologists), found that radiologists covered less distance in radiographs with diseases. For these radiographs, radiologists generally also looked for a shorter period of time at the image, made fewer fixations and saccades and fixated sooner on AOIs. For radiographs without diseases, radiologists covered higher distances than dental students. These results only partly support global processing of radiologists. While shorter process times of the image, fewer fixations, and sooner fixations on AOIs fit to fast global processing, fewer saccades and less coverage for radiographs with disease are not characteristics of global processing.

In contrast to the results of the aforementioned studies, another study examining eye movements on panoramic radiographs found no differences for eye movements on anomalies (time to first fixation, number and time of fixation and revisits) between experienced and inexperienced orthodontists (Bahaziq et al., 2019). The experienced orthodontists showed longer viewing times on the radiographs than the inexperienced orthodontist which is also in contrast to the above studies. Since the two groups also did not differ in their diagnostic performance, the similar results for eye movements on anomalies could be due to the slight difference between the expertise levels of the groups.

In summary, these few studies indicate that measures such as the first fixation on AOIs, fixation duration, number of fixations, scanning patterns and gaze coverage give insight into processes that differentiate expertise levels for dental medical image processing. Thereby, global processing also seems to play an important role for dental radiographs. However, the studies give no information on relevant gaze measures for lower levels of visual expertise development in dental students.

Two current studies examined visual search in dental students. Castner et al. (2018) showed that scan paths of dental students evaluating panoramic radiographs differ between a level before and after learning how to evaluate panoramic radiographs. Scan paths were not clearly differentiable in dental students in higher semesters after initial learning. Regarding the gaze measures mentioned above, Richter et al. (2020) found that the gaze behavior of dental students before and after learning changed in the following direction: students fixated AOIs

(anomalies with low prevalence) sooner, longer and more frequently. Besides, students showed a higher coverage after learning than before. On the whole, gaze behavior associated with the development of visual expertise in dental students is reflected in the time to first fixation on AOIs, the fixation duration on AOIs, the number of fixations on AOIs, and overall scan path and coverage. Thus, the relevant eye tracking measures appears to be very similar to those of other domains such as chest radiographs.

So far, however, no uniform and complete picture has emerged from the results of the studies with dental radiographs on the expression of specific measures and their relationship to different levels of expertise. In addition, the studies rarely recorded diagnostic performance, and the relationship of eye movement measures in visual search and diagnostic performance is still unclear and needs further investigations.

4.1.3 Cognitive skills and architecture in medical image processing

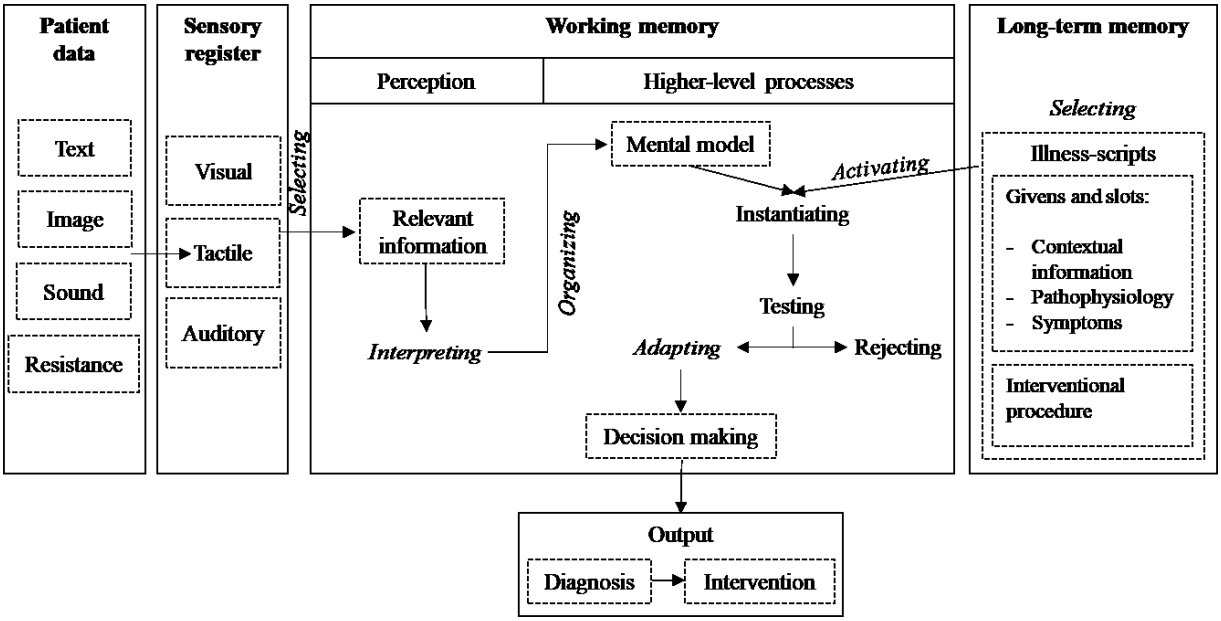
The evidence supports the global-focal models, which focus on the perception process of visual search in medical images. The following model extends this view with a stronger focus on the decision-making processes involved in medical image processing. The model of cognitive skills operating in medical diagnosis (Jarodzka, Boshuizen, et al., 2013) was combined with a model of Lesgold et al. (1988) and Boshuizen and Schmidt (2008). The model specifies cognitive skills and cognitive architecture that are involved in arriving at a medical diagnosis (see Figure 2). Here, the process of decision making is described more generally for medical diagnosis and not only for medical image diagnosis.

In the following, the model is discussed in more detail: As a starting point, the physician is faced with the patient's data in the form of text, images, sound or resistance (Jarodzka, Boshuizen, et al., 2013). These data enter working memory as visual, tactile, or auditory stimuli through a sensory register that is very limited in time but has an unlimited capacity (Baddeley, 1992). All conscious processes of information processing are located in the working memory, which is limited in capacity and time. Processing takes places in two areas (Jarodzka, Boshuizen, et al., 2013): First, during perception, the physician pays attention to certain elements and selects this from the sensory register. The physician must search for relevant information and distinguish it from irrelevant information. This interactive process between sensory register and working memory can already lead to pattern recognition and initial interpretations. Second, in the higher-level processes, the physician organizes the relevant information into a mental model. This mental model is created by activating and selecting illness scripts (see chapter 2.1) from long-term memory and is checked for validity in the next

step. This can result in either a rejection of the illness script or an adaptation to the current task. If the illness script is rejected an alternative script is selected. If the illness script fits the mental model and can be applied, a decision is made that results in the output of a diagnosis and interventions.

Figure 2

Model of cognitive skills operating in medical diagnosis (Jarodzka, Boshuizen, et al., 2013, p. 73)



Note. Reprinted from Catheter-based cardiovascular interventions: a knowledge-based approach, Jarodzka, Boshuizen, et al., *Cognitive skills in medicine,* p. 73, Copyright (2013), with permission from Springer Nature.

To date, the model as a whole has not been adequately tested for its validity for medical images. However, there is evidence to support certain parts of the model. Myles-Worsley et al. (1988) investigated the memory for chest radiographs of observers with different levels of expertise. Whereas recognition for chest radiographs with anomalies increased with higher expertise, recognition for normal chest radiographs without anomalies decreased. Thus, selective processing of relevant information as anomalies, as mentioned in the perception process in working memory of the model of cognitive skills (Jarodzka, Boshuizen, et al., 2013), seems to be associated with visual expertise for radiographs. Since expert selective processing and pattern recognition are also part of the global-focal models (see above), the empirical findings for these models also support the perceptual process of the model of cognitive skills.

Also for the second, high-level process which interacts with illness scripts from long-term memory, a study found this evidence: Experts used lower magnifications when diagnosing microscopic slides and could verbalize their observations into diagnoses, whereas intermediates used high magnifications and needed more time to reach diagnoses by checking more areas of the slide (Jaarsma et al., 2015). This shows that experts use encapsulation of illness-scripts to be efficient as proclaimed in the high-level process of the model. However, further evidence is needed to verify the model of cognitive skills in medical diagnosis for medical image processing.

4.1.4 Prerequisites for the processing of medical images

After considering the cognitive processes involved in image processing with the last model, a theoretical framework that highlights the skills and knowledge needed to successfully process medical images is presented.

Van der Gijp et al. (2014) developed a framework of knowledge and skills which trainees need to interpret medical images (radiographs). The knowledge and skills are assigned to three components in medical image interpretation - Perception, Synthesis, and Analysis. The Perception component consists of skills that are necessary to identify anomalies in radiographs, such as the use of efficient search strategies, the ability to discriminate normal from abnormal findings and pattern recognition. The synthesis component contains skills that trainees need to summarize multiple findings into a diagnosis and apply further treatment including information retrieval, connecting several findings, creating a diagnosis and deciding about treatments. The Analysis component consists of skills that are relevant for examining the features of anomalies, for instance, characterizing anomalies, comparing them with previous anomalies and discriminating relevant from irrelevant findings. Relevant to all three processes (Perception, Synthesis and Analysis) are knowledge of anatomy, pathology, radiological image techniques and clinical information/context as well as spatial abilities and image manipulation skills.

So far, the whole framework has not been validated. However, some aspects of the framework have been investigated in a review paper examining factors that influence visual search of radiologists (Ganesan et al., 2018). One factor that influences visual search was prior knowledge of clinical history which fits to the knowledge of clinical information/context of the framework. As another expertise factor, Ganesan et al. (2018) name comparative scanning strategies, which means that radiologists compare conspicuous features with other regions, for example the same anatomical region on the left and right side. These efficient search strategies are also stated in the framework of van der Gijp et al. (2014).

Further research is needed to evaluate this framework and to see if it can be useful for developing training that builds visual search and diagnostic skills to avoid diagnostic errors in medical image processing.

4.2 Diagnostic errors in medical image processing

After looking at how medical image processing works, this section highlights the errors that can occur. When physicians make mistakes in diagnosing medical images, this can have serious consequences for patients. For example, overlooking tumors or calcifications of the carotid artery, which can possibly lead to strokes, in panoramic dental radiographs can end badly for the patient (Constantine et al., 2018; Friedlander & Freymiller, 2003). The following section provides details on the diagnostic errors that can occur in the interpretation of radiographs, their types, and reasons.

High diagnostic error rates of missed anomalies in chest radiographs were first reported by Garland (1949). High error rates have also been found in many other radiologic image domains (cf. Waite et al., 2020), such as dental radiographs. Dentists missed 41% of anomalies in radiographs when searching for bony pathologies (Stheeman et al., 1996). In general, physicians make different errors in radiology (Pinto & Brunese, 2010): Errors occur during observation, for example resulting from a lack of alertness, distracting factors or characteristics of the anomalies. Further errors can be made in the interpretation which are influenced, for example, by the clinical history of the patient or the initial suspicion of a disease. Additionally, physicians may fail to propose an appropriate further procedure or may not appropriately communicate findings from radiographs to colleagues.

This dissertation focuses on errors which occur during the perception/observation process of evaluating radiographs. Two different kinds of errors may occur when observing and evaluating radiographs referring to signal detection theory (Green & Swets, 1988). False positive errors occur when physicians classify an area as abnormal although it is not. False negative errors refer to the cases where anomalies were not correctly identified as such. According to Waite et al. (2020), false negatives are the most important errors and occur frequently. Compared to false positive errors, false negatives cannot be revised in the further diagnostic or treatment process as patients appear healthy, which explains the error type's importance. In a study, radiology errors in 558 cases (radiographs or computed tomography [CT] scans for various body parts) in discrepancy meetings at a hospital were examined over

seven years (Donald & Barnard, 2012). The results showed that 80% of the errors were due to the perception process.

These false negative errors which result during the perception process have been further classified by eye tracking measures into search, recognition and decision errors (Kundel et al., 1978; Wu & Wolfe, 2019). Detection errors (also known as search errors) occur when an observer has not observed an anomaly and result from bottom-up processes. There are several factors that can lead to detection errors: for example, low prevalence of anomalies can affect diagnostic performance, and lead to premature search termination (Brunyé et al., 2019). Also, a satisfaction to search could cause detection errors (Brunyé et al., 2019). This means that when physicians find an anomaly in a radiograph, the probability of finding another anomaly decreases. This is because physicians are satisfied with finding one anomaly and thus end their search too quickly. However, this explanation does not seem to hold up (cf. Wu & Wolfe, 2019). Another factor is the depletion of working memory resources. In particular, when physicians are searching for multiple anomalies an overload of the working memory can lead to detection errors by not considering anomalies with low prevalence or low salience (Brunyé et al., 2019).

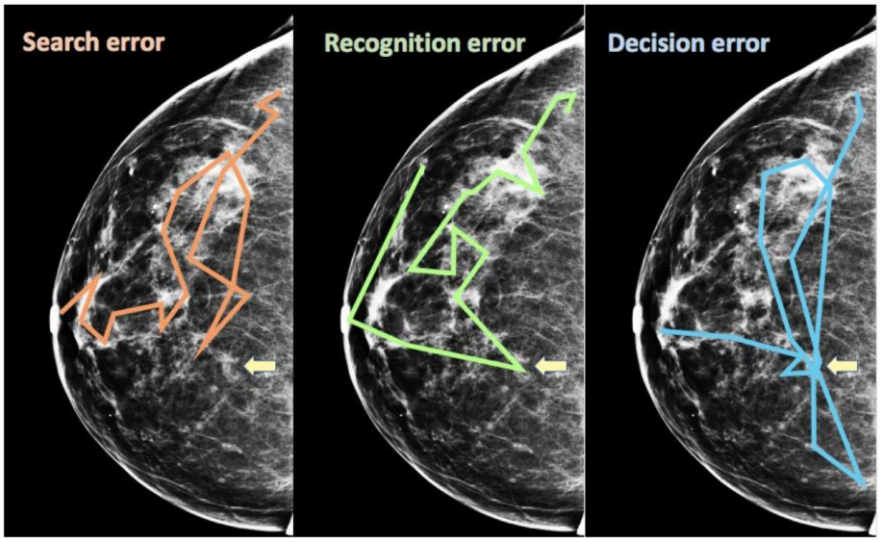
If the observer does not recognize relevant features of the anomaly, recognition errors occur (Brunyé et al., 2019). Reasons for this are, for example, lack of knowledge about patterns of anomalies or anatomical structures. If the observer looks at an anomaly, recognizes the features as suspicious, but ultimately decides against their relevance, this is called a decision error. Recognition and decision errors are mainly due to top-down processes: The observer perceives an anomaly visually but does not consider it further in the diagnostic process. The types of errors can be represented with a visual scan path of an observer in mammography (see Figure 3, cf. Wu & Wolfe, 2019). In Figure 3, the arrow points to the suspicious area with the anomaly and the colored line represents the eye movement in the form of a scan path. While for detection errors the scan path does not intersect the anomaly, this is the case for detection and decision errors. In decision errors, the anomaly was touched more often by the scan path with probably longer duration.

Kundel et al. (1978) originally defined the error types of missed anomalies using thresholds of gaze behavior. Observers did not look at the anomaly for detection errors, looked at anomalies for less than 1000ms for recognition errors, and looked at anomalies for more than 1000ms for decision errors (Al-Moteri et al., 2017; Donovan & Litchfield, 2013). However, this definition using a threshold is also problematic. First, the threshold is somewhat arbitrary. Second, there are context-specific differences. The fixation duration differed for chest x-ray and mammography radiographs with 200ms (cf. Al-Moteri et al., 2017). Thus, the threshold of

1000ms may not be appropriate for every type of radiograph or may be higher for anomalies that are more difficult to detect than others. Moreover, Brunyé et al., (2019) stated that the underlying assumption of an association between increased fixation time and successful detection of anomalies does not seem to hold. I have combined recognition and decision errors in my studies because they share the commonality that observers look at anomalies in both errors and both result from top-down processes. They are called top-down errors and detection errors are labeled as bottom-up errors.

Figure 3

Classification of false negative errors into search, recognition and decision errors by eye movements (Wu & Wolfe, 2019, p. 8)



Note. Reprinted from Vision, Wu & Wolfe, *Eye movements in medical image perception: A selective review of past, present and future,* p. 8, Copyright (2019) by Vision.

Kundel et al. (1978) were the first to investigate the frequency of these different error types in chest radiographs with eye tracking. They found that experts made about 30% detection errors, 25% recognition errors, and 45% decision errors. These results were based on the data from four experts. Donovan and Litchfield (2013) also investigated the different error types of chest radiographs with ten participants in four groups of different expertise level. They found no systematic differences in the frequency of error types among the different expertise groups. However, naïve observers made more detection errors than the other groups. In total, they found 12-22% detection errors, 29-48% recognition errors and 35-54% decision errors. A study by Manning, Ethell and Donovan (2004) also showed that 35% of the errors in the search for nodules in chest radiographs are detection errors, whereas 65% are recognition or decision

errors. Overall, the results of the aforementioned studies showed more decision errors than others and all types of errors were represented with a substantial number (Donovan & Litchfield, 2013; Kundel et al., 1978; Manning et al., 2004). However, the studies used only a relatively small number of participants and the values are difficult to compare due to different calculations of the error types.

As mentioned above, different radiographs with different characteristics influence visual search and the frequency of different error types (cf. Brunyé et al., 2019; cf. Gegenfurtner et al., 2011). So far, there has been no study investigating different error types in dental medicine radiographs. Thus, we do not know the frequency of different error types in panoramic radiographs. However, it could be assumed that the frequency differs from the frequency of chest radiographs due to different characteristics of the radiographs mentioned in section 4.1.2. Little is known about the distribution of error types in OPTs, and there are reasonable assumptions that they differ from chest radiographs. It is important that the error types are also studied for OPTs, as the study of the different types of errors provides detailed insight into the cognitive processes of physicians and the underlying problems in evaluating radiographs. Therefore, this insight is important to develop and apply appropriate interventions and training.

5. Training methods: Support for medical image processing

When medical students are first asked to interpret medical images, they still have a lot to learn in order to diagnose them correctly, such as how to recognize diseases and what strategies they should use to do so. This chapter explores what methods and interventions are used and can support medical image processing.

Evaluations of university curricula showed that whole curricula with multiple elements have positive effects on diagnostic performance (e.g. Manning et al., 2006; Richter et al., 2020). At universities in Europe, educators teach radiograph interpretation in very many ways and the undergraduate radiological curricula are not uniform (Kourdioukova et al., 2011). Moreover, "[i]t is still unclear which techniques make complete programs effective, and our educator is left with only a shallow understanding of what makes a specific instructional technique effective" (Kok et al., 2017, p. 4). Studies that evaluate specific instructional techniques are rare. Such previous studies found, for example, that instruction before practice is more effective than vice versa (Geel et al., 2018) or that testing outperforms studying radiographs (Baghdady et al., 2014). The following chapter summarizes the findings gained so far regarding training methods and begins with an overview of the different types of training.

5.1 Classifications of training for medical image processing

One way to classify training methods at a general level is provided in a review by Kok et al. (2017) who summarized 81 studies that examined instructional design in medical image processing. These studies could be classified according to the structure of the intervention: Most studies evaluated e-learning modules for medical image interpretation, followed by specific intervention techniques and complete curricula or courses. Only the investigation of specific intervention techniques appears to be very helpful for educators. The review indicates that teaching reasoning strategies and building cognitive schemata with, for example, concept-maps or through comparisons are important aspects of learning medical image interpretation. However, the authors mention that further studies are needed to reliably identify effective techniques. In particular, studies which consider research from visual expertise and individualized or self-regulated learning are lacking.

According to an approach of Kramer et al. (2019), visual search training in different domains, such as images of airport scanners, beach monitoring or medical images, can be divided into three types based on the target objectives to train: (1) training for the use of

technology and equipment, (2) training of object identification or (3) training of search strategies. Since this dissertation focuses on the personal ability to detect anomalies in medical images, only the classification of (2) and (3) is of further interest for this work.

The training of object identification (2) is designed to ensure that observers can identify targets such as anomalies based on their specific visual characteristics (cf. Kramer et al., 2019). This also requires background knowledge of the anomaly's characteristics, such as visual features, prevalence or pathology, and the specifics of the image, such as knowledge of anatomical structures (cf. van der Gijp et al., 2014). Thus, object identification training should also contain training about background knowledge. Object identification training appears to be most appropriate for specific targets (e.g., nodules in the lung) but is less effective for varied and dissimilar targets (cf. Kramer et al., 2019). As mentioned in chapter 4.1.2, OPTs may reveal a variety of anomalies in different locations. Thus, the effectiveness of object identification training might be limited for OPTs. Besides, object identification training appears to be most effective for short-term goals, e.g. when it is necessary to teach quickly what anomaly to look for (Kramer et al., 2019).

In contrast, the training of search strategies (3) is said to provide a longer-term change in visual search. The training of search strategies entails two components. First, training on what to fixate on, with systematic viewing aimed at higher coverage of images. However, it is not clear whether this also leads to better a detection of anomalies (cf. Kramer et al., 2019). Second, training for improved decisions which aims at observers extracting the important information from the image and developing decision skills. However, research on these aspects of training are lacking. This classification of training will guide the following review of studies on training in radiological images.

In the following sections, an overview of studies that investigated specific intervention techniques in radiograph interpretation based on the classification of Kramer et al. (2019) is provided. The studies on training for realistic medical images mentioned in the following two sections are summarized in Table 1.

5.2 Training of object identification

Identifying objects such as anomalies requires knowledge of the characteristic features of anomalies and background information on medical images, such as anatomical structures, or on the underlying mechanism of a disease (Kramer et al., 2019; van der Gijp et al., 2014). The following presents two methods that can be helpful in building a knowledge structure that

facilitates knowledge retrieval through teaching basic science or that supports categorization of anomalies through comparisons.

Some training methods that aim to support object recognition are methods which teach characteristic features of anomalies with verbal explanations (Baghdady et al., 2009, 2013). It turned out that it is better to learn these characteristic features together with basic science than with a structural algorithm that is supposed to be helpful in detecting anomalies, or only with a simple list of features without further information (Baghdady et al., 2009). When students learned the basic science mechanism integrated with the anomaly features (each feature along with its basic research mechanism), they later performed better diagnostically than students who learned the basic science mechanisms separately from the anomaly features (learning first a block of basic research followed by a block of features) (Baghdady et al., 2013). Thus, background knowledge as basic science seems to play a crucial role in identifying anomalies and the format of verbal explanations could also be helpful.

Another training method for medical images is compare-and-contrast training. Typically, students see two different medical images and are instructed to compare them. A web-based training tool COMPARE/Radiology from the University of Erlangen-Nuremberg uses this technique to help students to learn how to interpret radiographs (Grunewald et al., 2003). Comparisons support discrimination of relevant features and category learning (Hammer et al., 2008, 2009). Comparisons of same case examples are generally helpful for categorization tasks, whereas comparisons of different cases appeared to be not as beneficial on their own and should be guided, for example, by experts or instructions (cf. Hammer et al., 2008).

The following two studies show that comparisons of different cases, namely comparisons with a standard/normal case, can nevertheless be effective in the content of medical images. A study investigated how comparisons affect the detection of anomalies in depictions of skeletons (Kurtz & Gentner, 2013). Students who compared the depictions to a standard depiction were more accurate in finding anomalies in same and new depictions than students who studied twice as many depictions. In realistic medical images, comparison of chest radiographs with and without disease also positively affected diagnostic performance of medical students (Kok et al., 2013). However, these supportive effects occurred only for diseases located in a specific area, whereas no beneficial effects were observed for diffuse diseases that appeared in multiple areas of the radiographs. As the control group in this study compared radiographs of the same diseases instead of normal/disease comparisons, the above assumption of Hammer et al. (2008) does not seem to apply to medical images. Instead, the comparison of normal and disease radiographs seems to facilitate the discrimination of disease in specific areas. A second study showed that

also comparisons with two radiographs of same disease or of different diseases were efficient (Kok et al., 2015). Students who compared the same diseases were especially efficient in localizing anomalies and students who compared different diseases were efficient in discriminating anomalies. However, the comparison training in this study did not lead to an overall better detection of anomalies. One possible reason for this could be that further instructional support as mentioned above (cf. Hammer et al., 2008) is needed to make comparison training not only efficient but also effective.

In general, these studies indicate that comparison training is beneficial for students to learn how to discriminate and localize anomalies and therefore belongs to training that addresses object identification (cf. Kramer et al., 2019). To date, the results are not entirely clear and further studies are needed to verify the effect of comparison training and to investigate whether additional instructional techniques should be added to enhance the supportive effect. On the whole, training that focuses on object/anomaly identification in medical images seems to be helpful in medical image processing although more research is needed here.

5.3 Training of search strategies

One training method used to avoid detection errors is systematic viewing and is part of the training of search strategies (cf. Kok et al., 2016). The assumption is that for students who typically use a search-to-find method (Kundel et al., 2007) and show, for example, less frequent fixations on relevant areas (Jaarsma et al., 2014), this method helps to achieve a more complete visual coverage of the image. This method also plays a crucial role in university teaching and is mentioned in teaching literature: For trauma radiographs, for instance, the ABCs Image Interpretation Search strategy states key points for the procedure of diagnosing radiographs (Williams, 2013). The key points contain a check for adequacy, for different anatomic structures, and for full coverage of the radiograph. A systematic approach for OPTs is based on the division of the radiograph into four different regions of interest (Pasler, 1991), which should all be inspected.

So far, only few studies investigated this specific training method. One study reported that observers with poor diagnostic performance also did not systematically look at the radiographs (cf. Bahaziq et al., 2019). An experimental study by Kok et al. (2016) investigated this method using systematic viewing training, in which students were asked to adhere to a specific sequence of anatomical areas, and visual full-coverage training. In the visual coverage training, the students were asked to mentally divide the radiograph into segments and search separately in every segment. However, the study did not find improvements of diagnostic

performance in the training conditions compared to a control condition. Students who were trained with full-coverage training performed even worse than students in the other two conditions. A reason for this might be that the applied method of full-coverage training was a rather artificial way of obtaining full coverage and possibly interrupts the observers in their usual way of searching for anomalies. The application of this new method could tie up cognitive capacities of the observer that cannot be used to find anomalies. At least, it was found that students who used more systematic visual search also covered more areas of the chest radiographs. Similar to these results, systematic viewing did not lead to better diagnostic performance than nonsystematic viewing (van Geel et al., 2017). While in this study students in the systematic viewing group also viewed the radiographs more systematically, a positive relationship with the coverage of the radiograph was not found.

These two studies indicate that systematic viewing or full-coverage training are not particularly effective in this form. Perhaps this is because these training methods prevented the students' initial search strategy, missing short-term effects of search strategy training as mentioned by Kramer et al. (2019) or because reducing recognition errors is not the primary problem students face when interpreting radiographs. In this sense, Kok and Jarodzka (2017b) come to the conclusion that training should teach what anomalies look like (object identification) rather than systematic viewing (search strategies). However, due to the small number of studies, no definitive conclusions can yet be drawn from the results.

Other methods that train visual search strategies emphasize the perceptual component of visual search by using eye movements. Eye movement visualizations provide a relatively new method to support medical image processing. The following studies in this section refer to static gaze visualization. Kundel et al. (1990) conducted one of the first studies with eye movement visualizations in the area of radiology. Observers diagnosed a chest radiograph and then viewed a version of that radiograph overlaid with a visualization of their eye movements in the form of circles from the first view. In addition, observers were instructed that overlooked anomalies are typically associated with high visual attention. This intervention method resulted in a 16% increase in diagnostic performance compared with the control condition, which involved a second look at the radiograph without visualization. This benefit is also supported in another study using static gaze feedback in form of overlaid scan paths and fixations onto the radiograph (Donovan et al., 2008). Visual search after feedback seemed to change more in students than in experts or naïve observers, which may indicate that gaze feedback is particularly helpful in the learning process. In these two studies, diagnostic performance improved for the images for

which the viewers had previously received gaze feedback. Thus, it remains to be seen whether gaze feedback can have a general effect on other images as well.

The underlying mechanism of this method seems to be a directing of attention to the relevant areas, which is also possible with other methods such as signaling (Richter et al., 2016; van Gog, 2014). Guiding observers' attention can also change cognition thereby influencing the performance of a task (Grant & Spivey, 2003). Contrary the above results, a study by Drew and Williams (2017) which investigated eye movement feedback in a controlled artificial setting (searching a target in a landscape image) found no positive effects on target detection. The authors cite the way gaze feedback is given as a reason for ineffectiveness. Gaze feedback was displayed in the form of colored rectangles superimposed on the image, which contrasts with the more "natural" display method in the above studies. It remains to be said that further studies are needed to make definitive statements about the effectiveness of the method.

One reason for this small number of studies in the field of radiology could be that the method has evolved to dynamic gaze visualizations of a model: Eye movement modeling examples (EMME). This intervention method consists of learners seeing videos of the eye movements of a model, usually experts, as they complete the task being learned (cf. Jarodzka et al., 2013). Thereby, EMMEs show the scan path of the models with fixations (and saccades) in temporal sequence. This gives the learner insight into the process, how the model performs the task, and the visual attention of the model. In addition to directing attention to relevant areas, which is also present in static gaze feedback (see previous section), EMME adds modeling that can additionally convey strategies, e.g., visual search strategies, and acts as an instructional tool. In worked examples, learners can gain insight into the strategies of experts through the modeling behavior and use the insight for themselves (cf. van Gog & Rummel, 2010). Some studies also simultaneously added verbal explanation to the display of eye movements (Gegenfurtner, Lehtinen, et al., 2017; Jarodzka et al., 2012; Jarodzka, van Gog, et al., 2013). These offer the advantage that the reasoning of the models can be explicitly verbalized, and learners do not have to capture them only via eye movements. Most of the studies used didactical verbal explanations which could make their reasoning more understandable for the learners (cf. Isaacs & Clark, 1987).

According to observational learning theory, learners are able to adopt a model's behavior (Bandura, 1971). Aside from the aforementioned attentional focus that can also be guided by a model, learners get ideas about how to combine and assemble the components of a task by observing the model's behavior on that task (Bandura, 1971). Applying this to the visual search of medical images, models can direct attentional focus to the relevant areas in the image while

teaching visual search strategies, for example left and right-side comparisons of anatomical areas. Furthermore, Bandura assumes that learners can also develop new, generalized behavior from the observed behavior of the model and thus transfer the behavior to another situation with similar conditions (Bandura, 1971). Accordingly, in our case, models should also be helpful, for example, when learners evaluate different medical images than the previous model.

The concept of cognitive apprenticeship also emphasizes the importance of models for learning (Collins & Kapur, 2014). The concept states that for acquiring cognitive skills apprenticeship with a focus on expert processes and learning in a context is beneficial. Various methods can be used to build expertise in an area, including modeling. Thereby, students observe the experts who is performing a task. The challenge in cognitive domains consists in making the internal processes and activities visible. Then, the students can build a model of the processes which is needed to perform the task. EMME provide an opportunity to make such internal processes of the expert's visual search on medical images visible.

EMME have first been applied in other domains. EMME were beneficial in multimedia learning and successfully supported learning performance and the previous processing of text and pictures (Mason et al., 2015, 2017; Scheiter et al., 2018). In the context of problem-solving and reasoning performance, EMME also showed positive effects (Jarodzka, van Gog, et al., 2013; Litchfield & Ball, 2011; van Marlen et al., 2018). However, single studies failed to find positive effects on learning for a problem-solving task (van Gog et al., 2009; van Marlen et al., 2016). In the medical domain, EMME supported clinical reasoning in students who learned to diagnose epileptic seizures in infants (Jarodzka et al., 2012). Thereby, students saw videos of infants' body movements with the superimposed gaze behavior of an expert.

Some studies have also applied EMME in medical image processing. Litchfield et al. (2010) investigated EMME in the context of searching for nodules in chest radiographs with three experiments. In the first experiment, EMME supported diagnostic performance regardless of model's expertise level. The second experiment investigated the modality of EMME and found that only task related EMME led to positive effects on performance, which were only constants for novices and not for more experienced observers. In the third experiment, the influence of modality of EMME (task related vs. not task related) and its positive effects on diagnostic performance were verified again. EMME also showed positive effects for interactive medical images (for Computer Tomography (CT) scans and Positron Emission Tomography (PET) scans: Gegenfurtner, Lehtinen, et al., 2017; for CT: Seppänen & Gegenfurtner, 2012). EMME of experts who performed an interpretation of a CT scan supported students' diagnostic performance (Seppänen & Gegenfurtner, 2012). Besides, the visual search also changed to a

more focused and selective search on the relevant areas. Gegenfurtner, Lehtinen, et al. (2017) also found that EMME positively affect diagnostic performance when the observer interpreted the same case as shown in the EMME video before. However, the EMME video only supported accuracy in different CT scans for experts but not for novices. The visual search seemed to change to a more focused search for both tasks, with the same or different CT scans, with more fixations on relevant areas and in general longer fixations in the relevant areas. On the whole, these studies provide first evidence that EMME may support medical image processing.

Since EMME also fall into the training category of search strategies, according to Kramer et al. (2019) they should mainly show long-term effects. In the above studies, which only investigated short-term effects, EMME also proved effective after a short period of time. In general, training of search strategies such as EMME and gaze feedback appears to promote medical image processing while it is not clear if systematic viewing is beneficial for medical students.

Table 1. Main characteristics of training studies to foster diagnostic performance in medical images

Authors	Medical images Participants (*N*)	Design and methods	Training	Effects
Baghdady et al. 2009	Dental radiographs Dental and dental hygiene students ($N = 96$)	Three groups with different learning strategies (basic science vs. structured algorithm vs. feature list)	Three different learning strategies/ training material: • Basic science (radiographic features of disease + information on basic disease mechanism) • Structured algorithm (radiographic features of disease + general algorithm to analyze lesions) • Feature list (radiographic features of disease)	**Diagnostic performance:** Basic science group better than others on immediate and delayed test
Baghdady et al. 2013	Dental radiographs Dental students ($N = 51$)	Two groups with different learning strategies (integrated basic science with clinical features vs. segregated basic science and clinical features)	Two different learning strategies: • Each radiologic feature of disease integrated with underlying disease mechanism (basic science) • First learning of disease mechanism and afterwards radiologic features	**Diagnostic performance:** Integrated basic science group better than segregated basic science group
Kok et al. 2013	Chest radiographs Medical students ($N = 61$)	Two groups: normal-disease comparison of radiographs vs. control group – studying radiographs of disease	Comparison of radiographs	**Diagnostic performance:** • Comparing radiographs led to better diagnosis of disease in specific area • No effects on diagnosis of diffuse disease (involving more areas)

Authors	Medical images Participants (*N*)	Design and methods	Training	Effects
Kok et al. 2015	Chest radiographs Medical students ($N = 48$)	Four groups with different comparison techniques (same-disease comparison vs. different-disease comparison vs. disease/normal comparison vs. no comparison/control condition)	Comparison of radiographs	**Diagnostic performance:** • Detection of anomalies was not affected by the interventions • Disease/normal comparison led to better identification of normal cases **Processing:** • Same-disease comparison led to higher efficiency in localizing anomalies • Different-disease comparison led to higher efficiency regarding discrimination of anomalies
Kok et al. 2016	Chest radiographs Medical students ($N = 75$)	Three between training conditions (systematic viewing vs. full coverage vs. non-systematic viewing)	Systematic viewing / full coverage training	**Diagnostic performance:** • Worse performance after full-coverage training than other training methods • No training benefits **Visual search:** • Positive correlations between systematic viewing and coverage of the image
Van Geel et al. 2017	Chest radiographs Medical students ($N = 60$)	2 x 2 mixed methods: Time as within factor (pre-, post-test), training type as between factor (systematic vs. nonsystematic viewing)	Systematic viewing	**Diagnostic performance:** • Higher sensitivity after training ($\eta_p^2 = .11$) • No effect of training type • No effect on specificity **Visual search:** Systematic viewing group showed higher systematicity ($\eta_p^2 = .29$) No effect on coverage

Authors	Medical images Participants (*N*)	Design and methods	Training	Effects
Kundel et al. 1990	Chest radiographs Radiology residents ($N = 6$)	Two between conditions (gaze feedback vs. second look) at follow-up test groups returned	Static gaze feedback	**Diagnostic performance:** • Positive effect of gaze feedback on accuracy
Donovan et al. 2008	Chest radiographs Participants with different expertise level ($N = 40$)	2 x 4 mixed methods: expertise as between factor (naïve vs. student-level 1 vs. student-level 2 vs. experts), gaze feedback as within factor (pre and post)	Static gaze feedback	**Diagnostic performance:** • Positive effect of gaze feedback (within group receiving feedback) **Visual search:** • Eye movements of naïve and expert observer were less affected by feedback than students (both levels)
Litchfield et al. 2010	Chest radiographs **Experiment 1** Participants with different expertise level ($N = 48$)	**Experiment 1** 2 x 2 x 3 mixed methods: expertise as between factor (novice vs. experienced), model expertise as between factor (novice vs. experts, viewing condition as within factor (free search; image preview; eye movement preview)	EMME video	**Experiment 1** ***Diagnostic performance:*** • Positive effect of eye movement preview for novices and experts regardless of model expertise ($\eta^2 = 0.41$) ***Processing:*** • With eye movements preview longer decision times than other viewing conditions ($\eta^2 = 0.61$)
	Experiment 2 Participants with different expertise level ($N = 60$)	**Experiment 2** 2 x 3 between design: factor expertise (novice vs. experienced), factor viewing condition (image preview vs. expert search preview vs. unrelated preview)		**Experiment 2** ***Diagnostic performance:*** • Positive effects only for expert's search preview ($\eta^2 = 0.24$) • Consistent improvement only for novices ($\eta^2 = 0.48$)
	Experiment 3 Novice radiographers ($N = 40$)	**Experiment 3** Four between viewing conditions (naïve-no-task vs. naïve-search vs. incongruent-search vs. expert-search)		**Experiment 3** ***Diagnostic performance:*** • Positive effects of experts-search and naïve-search eye movements ($\eta^2 = 0.42$)

Authors	Medical images Participants (*N*)	Design and methods	Training	Effects
Seppännen & Gegenfurtner 2012	CT visualizations Medical students ($N = 26$)	2 x 2 mixed method: intervention as between factor (intervention vs. control group), time as within factor (pre-, post-test)	EMME video	**Diagnostic performance:** • No difference between groups at pre-test • Intervention group improved in accuracy and sensitivity **Visual search:** • No differences between groups at pre-test • Intervention group looked more on relevant areas and less on redundant areas after intervention
Gegenfurtner, Lehtinen et al. 2017	Dynamic PET/CT visualizations Participants with different expertise level ($N = 23$)	3 x 3 mixed-methods: expertise as between factor (PET experts vs. CT experts vs. novices), case as within factor (baseline - before EMME vs. retention - after EMME same case vs. transfer – after EMME different case)	EMME video + think aloud protocol + screen action	**Diagnostic performance:** • Positive effect on accuracy in retention task for experts (Cohen's $d = 0.55$) and novices (Cohen's $d = 1.94$) • Positive effects on specificity in experts and novices ($\eta_p^2 = 0.956$) **Visual search:** • More fixations on task-relevant areas after training for experts (retention: Cohen's $d = 1.67$, transfer: Cohen's $d = 1.23$) and novices (retention: Cohen's $d = 0.70$, transfer: Cohen's $d = 0.73$) • Less fixations on task-redundant areas after training in retention task and compared to transfer task • Longer fixation durations after training for task-relevant and -redundant areas

6. Overview of research questions and studies

Diagnostic errors are very common in medical image interpretation, such as dental radiographs (Pinto & Brunese, 2010; Stheeman et al., 1996). To prevent dentists from committing diagnostic errors that can have serious consequences, effective training methods are needed, especially for dental students. So far, only few evaluated training methods exist for medical image processing (Kok et al., 2017). Specifically, in the domain of dental panoramic radiographs (OPT), to my knowledge, no specific intervention techniques have been previously studied or evaluated. It is particularly important in this area to conduct research on training methods, as it cannot be assumed that training methods used for other medical images, e.g. chest radiographs, will have the same effect here due to the different image characteristics (cf. Gegenfurtner et al., 2011). These image characteristics, such as the higher number of anomalies in OPTs, could have implications for the underlying psychological processes of visual search and object recognition and the probability of errors (Wu & Wolfe, 2019). The main research question is derived from the lack of evaluated training methods for OPT interpretation and the poor diagnostic performance: How to support OPT interpretation of dental students? This question can be further specified into the following: What methods are beneficial to promote anomaly detection and intensify visual search? To close the research gaps, the aim of my dissertation is to develop and evaluate specific training techniques for OPT processing in dental students. Besides, investigations of these training methods could also reveal conclusions about the underlying problems of OPT interpretation. I have developed three different training interventions and assume that the use of these interventions would lead to better detection of anomalies and more intensive visual processing of OPTs as measured by eye-tracking.

The resulting hypotheses for all studies are the following: The training interventions should improve the diagnostic performance of dental students (H1). Especially the detection of anomalies should improve, but also false positive errors and different error types were examined exploratorily. Based on the research of Richter et al. (2020), who studies the effects of training on visual processing of dental students, and the global-focal model (Nodine & Mello-Thoms, 2000), it is expected that the training interventions of this dissertation lead to higher visual coverage of OPTs (H2; not for Study 2 that only trained object identification but no global search strategies), an increased fixation time on anomalies (H3), an increased number of fixations on anomalies (H4), and a decreased time to first fixation on anomalies (H5). The specific hypotheses for the groups of participants and addressed anomalies are described in the corresponding manuscripts.

In Study 1, I investigated individualized full coverage training designed to help dental students to search in all areas of the OPTs and thereby reduce the number of missed anomalies. In total, 61 dental students either received the training for five OPTs in the intervention group or diagnosed five OPTs in the control group in a pre- and post-test setting. The training consists of five OPT comparisons where the students simultaneously saw an OPT with two static eye movement visualizations: The visualization of a peer model who showed full coverage and the same OPT with their own eye movement visualizations, recorded in the pre-test. This implicit approach is to make students aware of their visual search and its defects, to realize that they have not looked at the peripheral areas with maxillary sinuses, for example, as opposed to the model, and to encourage them to cover the complete radiographs in the visual search. Unlike other full-coverage or systematic viewing training which did not improve diagnostic performance (Kok et al., 2016; van Geel et al., 2017), this method does not artificially interrupt the visual search in the radiograph, but instead shows the natural and individual visual search behavior for participants to reflect on. It was expected that the training addresses anomalies located in the neck, maxillary sinuses, jawbone and jaw joint (hereinafter referred to as peripheral anomalies) with lower prevalence (Constantine et al., 2018; Vallo et al., 2010) more than anomalies located in the oral cavity (hereinafter referred to as central anomalies) which are more common.

In Study 2, the training addressed the larger part of the problem, recognition and decision errors which could be inferred from the results of Study 1. Two training sessions were evaluated simultaneously that addressed either the recognition of peripheral anomalies or central anomalies with a crossed over design. 78 dental students participated in this study and were assigned to two groups. One group received first the training to recognize peripheral anomalies and second the training to recognize central anomalies, while the other group received the training in the reversed order. The training contained comparisons of two OPTs with and without disease and comparisons of two OPTs with the same disease. Colored highlights of relevant anatomical structures or pathological areas could be superimposed on the comparisons. In addition, a verbal description of the visual characteristics of the anomalies was provided to support their identification. Here, the training of this study aimed at object identification by providing information about the characteristic features and focusing on their discrimination, rather than teaching a search strategy as in the first study.

For Study 3, training that combines search strategy and object identification was developed. In total, 86 dental students performed a pre- and post-test divided into intervention or control groups. In the intervention group, students saw three EMME videos of experts with

didactical verbal instructions between the tests. While students could adopt visual search strategies from the eye movements of the experts in the EMME video, the concurrent verbal instructions were intended to help more in object identification. The evaluation of the training was replicated in an online study of 31 dental students without measuring eye movements.

7. Study 1:

How to support dental students in reading radiographs: Effects of a gaze-based compare-and-contrast intervention

Published as: Eder, T. F., Richter, J., Scheiter, K., Keutel, C., Castner, N., Kasneci, E., & Huettig, F. (2021). How to support dental students in reading radiographs: effects of a gaze-based compare-and-contrast intervention. *Advances in Health Sciences Education,* 26(1), 159-181. https://dx.doi.org/10.1007/s10459-020-09975-w

7.1 Abstract

In dental medicine, interpreting radiographs (i.e., orthopantomograms, OPTs) is an error-prone process, even in experts. Effective intervention methods are therefore needed to support students in improving their image reading skills for OPTs. To this end, we developed a compare-and-contrast intervention, which aimed at supporting students in achieving full coverage when visually inspecting OPTs and, consequently, obtaining a better diagnostic performance. The comparison entailed a static eye movement visualization (heat map) on an OPT showing full gaze coverage from a peer-model (other student) and another heat map showing a student's own gaze behavior. The intervention group (N = 38) compared five such heat map combinations, whereas the control group (N = 23) diagnosed five OPTs. Prior to the experimental variation (pre-test) and after it (post-test), students in both conditions searched for anomalies in OPTs while their gaze was recorded. Results showed that students in the intervention group covered more areas of the OPTs and looked less often and for a shorter amount of time at anomalies after the intervention. Furthermore, they fixated on low-prevalence anomalies earlier and high-prevalence anomalies later during the inspection. However, the students in the intervention group did not show any meaningful improvement in detection rate and made more false positive errors compared to the control group. Thus, the intervention guided visual attention but did not improve diagnostic performance substantially. Exploratory analyses indicated that further interventions should teach knowledge about anomalies rather than focusing on full coverage of radiographs.

7. 2 Introduction

Reading radiographs such as orthopantomograms, (OPTs, panoramic radiographs of the upper and lower mandible including dentition), is a standard diagnostic procedure in the daily work of dentists, but is an error-prone process (Stheeman et al., 1996). Undetected or misinterpreted anomalies can have serious consequences for patients. For instance, carotid calcifications in the soft tissues of the neck can potentially lead to a stroke (Friedlander et al., 2005; Tamura et al., 2005). To avoid these diagnostic errors, it is important to start training early. However, training targeted at improving students' diagnostic performance is lacking (Kok et al., 2017). We therefore designed this study to evaluate a gaze-based training intervention for dental students.

Diagnostic errors in medical image interpretation can be classified into two groups (Gegenfurtner et al., 2017). Diagnosing a feature as abnormal although it does not represent an anomaly corresponds to a false positive error, whereas diagnosing a feature as normal although it represents an anomaly corresponds to a false negative error. False negative errors are particularly problematic, as health-threating situations are overlooked while false positive errors can be corrected in the further course of an albeit unnecessary treatment. False negative errors can be further classified into detection, recognition and decision-making errors (Al-Moteri et al., 2017; Donovan and Litchfield, 2013; Kundel et al., 1978). Detection errors occur when an observer does not visually attend to, or overlook, an anomaly; these errors result from misguided perception processes (bottom-up process). Recognition errors occur when an observer attends to anomalies, but lacks knowledge about characteristic features of anomalies and healthy structures, thereby not recognizing the anomalies (top-down process). Decision-making errors occur when the observer fixates on the anomaly and recognizes ambiguous features, but decides against their clinical relevance (top-down process).

The frequency of these errors, which has been mostly investigated in chest radiographs using eye tracking (Donovan and Litchfield, 2013; Kundel et al., 1978; Manning et al., 2004) and using radiographs and computer tomography (CT) images (Donald and Barnard, 2012), differs. Donald and Barnard (2012) found 80% of errors were detection errors and the aforementioned eye-tracking studies showed a maximum of 35% detection errors. A possible explanation for this difference could be the types of images (radiographs and CTs), which are related to different anatomical areas. This explanation corresponds to findings of a meta-analysis by Gegenfurtner et al. (2011) suggesting that domain specificity and task characteristics influence visual search. Consequently, findings from studies conducted with a certain type of image and task cannot be directly transferred to other images and tasks.

To the best of our knowledge, only two other studies have investigated visual search in OPTs (Grünheid et al., 2013; Turgeon and Lam, 2016) and none of them investigated error types. There are good reasons to assume that the frequency of error types in OPTs differs from that in chest radiographs. Chest radiographs typically indicate no more than five anomalies, which is rather a small number compared to anomalies that can be found in OPTs (Donovan and Litchfield, 2013; Kundel et al., 1978). In the OPTs used in the present study, which were obtained from patients reporting no obvious complaints, there were up to 26 anomalies within one OPT. Consequently, the likelihood for detection errors is higher in OPTs due to the larger number of anomalies. In contrast with experts, dental students are likely to commit even more detection errors, because they need to apply a search-to-find method (Kundel et al., 2007;

Nodine and Mello-Thoms, 2000). Additionally, detection errors are more likely to occur for low-prevalence anomalies rather than high-prevalence anomalies. Low-prevalence anomalies are located more often in the periphery compared to the central areas of the oral cavity (Constantine et al., 2018; Vallo et al., 2010).

The different error types as well as visual search processes in image reading can be investigated by means of eye tracking (Kok and Jarodzka, 2017a), where the gaze of a person inspecting a stimulus is recorded with a camera. The gaze is later analyzed with respect to its spatial and temporal characteristics, thereby allowing statements about which elements of the stimulus were looked at, when, and for how long. In these analyses, the gaze is further divided into separate events, namely, fixations and saccades. During a fixation, the gaze remains focused on one area of an image and information about this area can be processed (Just and Carpenter, 1980; Kok and Jarodzka, 2017a). Saccades are fast movements to re-position the eye and hence change the focus of attention. For these two types of events, various different eye tracking measures can be determined that provide important insights into a person's gaze behavior. Commonly used eye tracking measures for visual search in medical images are the time to first fixation regarding a specific area of interest (AOI; e.g., an anomaly), total fixation time, the number of fixations, the number and length of saccades as well as image coverage (van der Gijp et al., 2017). The time to first fixation denotes the time it takes a person to first attend to an anomaly. Number of fixations and fixation time (duration of fixations) on AOIs typically reflect more intense processing of this area. The coverage denotes the degree to which a person has inspected an image by having fixated in multiple areas. In the present study, these measures were used to investigate the effectiveness of the intervention, thereby assuming that the intervention would improve diagnostic performance via changing students' visual search behavior.

So far, there is little research describing and evaluating training approaches for improving visual interpretation of radiographs (Kok et al., 2017). Nevertheless, some systematic approaches for interpreting radiographs do exist. A systematic approach for OPTs is based on the division of the radiograph into four different regions of interest (Pasler, 1991), which should all be inspected to prevent students from missing anomalies. Especially novices, who typically process only small parts of images (Jaarsma et al., 2014), should use a full coverage approach in order to detect all anomalies. Previous research has, however, shown that a full coverage training may not necessarily lead to better diagnostic performance. In a study by Kok et al. (2016) medical students were asked to mentally divide a radiograph into segments and then separately search in every segment, which did not yield better diagnostic performance. An

explanation could be that the training was rather artificial, and possibly interrupted the students in their own strategies of searching for anomalies.

An innovative instructional method to enhance full image coverage is to illustrate adequate visual search behavior by showing how a role model (e.g., an expert or advanced learner) would perform these search processes. Here, eye tracking is not only used for measuring attentional processes, but also as an instructional tool (cf. Scheiter & Eitel, 2016). Gaze-based modeling has been used effectively in various contexts to support learning (e.g., multimedia learning: Mason et al., 2015; clinical reasoning: Jarodzka et al., 2012). When applying eye movement modeling to diagnostic search tasks, the gaze behavior of a person (i.e., the model) searching for anomalies is visualized and displayed as training material to learners. The learners observe the model's gaze behavior and are supposed to incorporate his/her behavior into their own repertoire of cognitive strategies (van Merriënboer & Kirschner, 2007). Eye movement modeling has been shown to foster diagnostic performance (for chest radiographs: Litchfield et al., 2010; for PET/CT: Gegenfurtner, Lehtinen, et al., 2017). Against this backdrop, we used a model's gaze to visualize full gaze coverage of a radiograph, which is expected to improve coverage – and in turn diagnostic performance - in students. We used static gaze visualizations (i.e., heat maps) where the model's distribution of visual attention was visualized and superimposed onto the OPT. Thus, areas attended by the model were highlighted while the underlying structure and the rest of the image remained visible (cf. Jarodzka et al., 2012).

More important, not every model is equally helpful. The model-observer similarity effect states that learners are more likely to adopt the model's behavior if s/he is perceived as being similar (Schunk, 1987; Schunk & Hanson, 1985). Accordingly, Krebs et al. (2019) found that students with low prior knowledge profited from eye movement modeling only if the models were introduced as peer-models but not as expert-models. Moreover, radiologist experts use search strategies that cannot be deployed by novices yet (i.e., global-focal search; Kundel et al., 2007; Nodine and Mello-Thoms, 2000), who lack the necessary knowledge. In particular, experts require a quick glance at a suspicious area of an image only, whereas good performance in students is likely to be characterized by intense processing of all areas of an image. Thus, it is questionable whether students could learn from gaze visualizations obtained from an expert model (cf. van der Gijp et al., 2017). Therefore, we chose heat maps from other, more advanced students who showed full coverage of the OPTs and intense processing of all its areas as peer-models to guarantee a high model-observer similarity.

An approach that combines modeling with individualized learning is the compare-and-contrast approach (van Merriënboer and Kirschner, 2007). Kok et al. (2013) showed that students who compared and contrasted chest radiographs indicating diseases against radiographs without diseases improved their diagnostic skills compared to students who only studied radiographs indicating diseases. Against this backdrop, in the present study we asked students to compare and contrast the gaze coverage of a peer-model with a gaze display of their own that had been recorded in an earlier trial to encourage more active processing of the model's gaze display and to enhance the students' understanding of systematic search.

7.2.1 The present study

The goal of the study was to improve dental students' diagnostic performance of reading OPTs by encouraging them to fully cover the image during visual inspection by means of a training. A full coverage of an OPT should help to avoid overlooking peripheral anomalies and thus reduce the number of detection errors. To support a full coverage, we combined two different instructional approaches within a gaze-based intervention. First, we presented students with a static gaze visualization obtained from a peer learner adjunct to their own gaze visualization. The peer-model's gaze visualization served as reference standard to which the participants could compare their own search behavior. Second, we asked them to compare and contrast the two visualizations. The visualizations were heat maps, where more saturated colors indicated more attention to an area. The intervention group was contrasted with data from a business-as-usual control group, who took part only in the routine training offered to the dental students.

First, we hypothesized that the compare-and-contrast modeling intervention leads to a more complete visual search, which should be reflected in a more comprehensive coverage when inspecting radiographs. Thus, the coverage should increase in the intervention group from pre- to post-test, whereas the coverage in the control group should not change over time (Hypothesis 1).

Second, we expected the change in gaze behavior due to training to differ between anomalies located in peripheral areas and those in central areas (Hypotheses 2a-c). Consequently, we assumed a three-way interaction between time (pre- vs. post-test), intervention (intervention vs. control group), and location (peripheral vs. central). The number of fixations (Hypothesis 2a) and the fixation time (Hypothesis 2b) for peripheral anomalies should increase from the pre- to the post-test in the intervention group, but not in the control group. Additionally, we assumed that students in the intervention group, but not in the control

group, would fixate on anomalies in the peripheral area in the post-test sooner than in the pre-test (Hypothesis 2c). No changes were expected for central anomalies for any of the gaze measures.

Because students in the intervention group were expected to show improved visual coverage of the OPTs, it was also assumed that they would conduct fewer detection errors, resulting in better diagnostic performance. Thus, the training should improve diagnostic performance from pre- to post-test as a function of anomaly location (three-way interaction: time x location x intervention). The diagnostic performance in the intervention group, but not in the control group, should increase from the pre- to the post-test especially for peripheral anomalies; fewer, if any, improvements were expected for central anomalies (Hypothesis 3).

7.3 Methods

7.3.1 Participants and design

78 dental students, who were either in their 7th or 9th semester, participated voluntarily in the experiment. At the Dental Medical School of the University of Tübingen, all dental students are requested to take part in a radiology course, where they are taught about radiation, imaging techniques, and radiograph interpretation in the 6th semester; this course includes massed practice of interpreting 100 images, mostly OPTs (Richter et al., 2020). They graduate after the 10th semester. On average there are 22 students in each study cohort, with a new cohort starting each summer and winter term. Accordingly, when inviting 7th and 9th semester students in two consecutive terms to participate in the study, a full-scale survey would have contained 88 students – which was nearly achieved. Accordingly, our sample was reasonably representative of the overall population of dental medical students at these two study levels. As incentives, students received a 15€ book voucher and individual feedback regarding their performance and gaze behavior at the end of the semester. 14 students did not complete the whole experiment (i.e., they did not participate in either the pre-test or in the post-test session) and thus had to be excluded from data analyses. Data from 3 students were excluded due to technical problems. The control group consisted of 23 students in the 9th semester (N = 23 (16 female); age = 25.39 years, SD = 2.48). The intervention group consisted of 7th (N = 23 (11 female); age = 24.27 years, SD = 2.61) and 9th (N= 15 (11 female); age = 27.52 years, SD = 3.13) semester students. Hence, students in the control group and intervention group came from different semesters. However, we know from previous data collections that there are no processing or performance differences between these two semesters (cf. Castner et al., 2018), which if anything, would

work against our hypotheses anyway. The data of the control group and the pre-tests of the intervention group were collected as part of a larger study where we investigated the development of visual expertise in a longitudinal study design involving all dentistry study semesters.

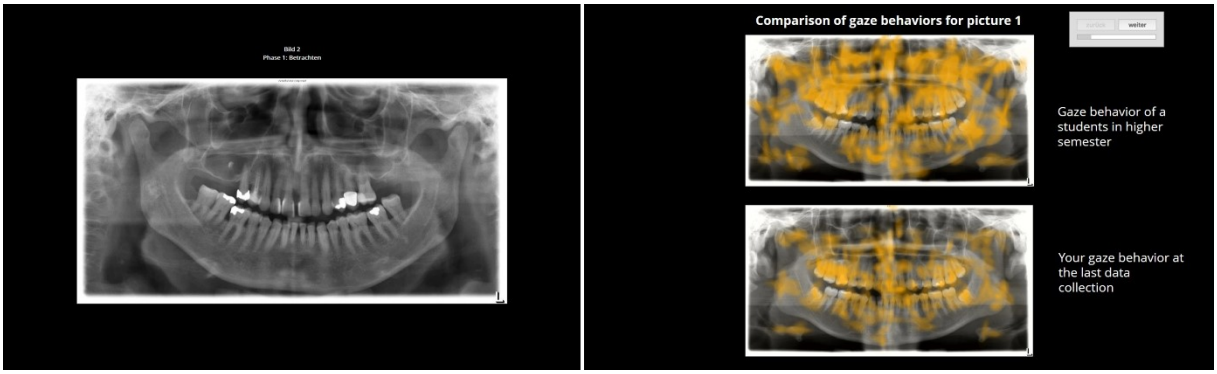
7.3.2 Materials and apparatus

OPTs

Overall, 20 OPTs that were recorded during routine checks in the university hospital were used as test and intervention stimuli. The OPTs were grouped into two sets of 10 OPTs each (set A and B). Set A was further separated into two sets A1 and A2 with five OPTs each. For the comparison between the control and intervention group, we used set A in the pre-test, set A1 and B in the post-test. Three OPTs showed no anomalies, whereas the other OPTs showed between 1 and 26 anomalies. Set A contains 79 central anomalies and 16 peripheral anomalies. Set B contains 41 central anomalies and 9 peripheral anomalies. Two experts (a maxillofacial radiologist and a prosthodontist both with over 13 years of clinical experience; co-authors of this paper) examined the OPTs and created solution templates. The OPTs had a sufficient clinical image quality (without positioning errors) and were displayed with a size between 1362 x 750 pixels and 1552 x 750 pixels (constant height for all OPTs) on a laptop (see Figure 4, left panel).

Figure 4

OPT displayed on a laptop (left panel) and compare-and-contrast training intervention for one OPT (right panel).



Training intervention

The compare-and-contrast training intervention contained heat map combinations for 5 OPTs. For every heat map combination, two heat maps were presented among each other (see

Figure 4, right panel). The upper heat map represented the gaze behavior of an advanced student (peer-model) searching for anomalies. The lower heat map showed the current participant's gaze behavior recorded during the pre-test. The heat map comparison had the title 'comparison of gaze behavior for picture 1'. The peer-model was labeled as the 'gaze behavior of an advanced student'. The participant's heat map was labeled by 'your gaze behavior at the last data collection'.

Heat Maps

The heat maps were constructed using the software Eyetrace (Kübler et al., 2015). The heat maps illustrated fixations and the duration of saccades where their location and intensity was displayed in orange (see Figure 4, right panel). Five individual heat maps showing students' individual gaze during OPT inspection in the pre-test from set A2 were generated for each student. If students had not participated in the pre-test or had low eye-tracking quality, their recordings obtained in previous session (approx. 2 months before the pre-test) were used to create the heat maps. The five peer-model heat maps were created for the same OPTs as those used for the individual heat maps. We selected them from students who showed a full coverage and intense processing of all relevant areas of the OPTs, especially for peripheral areas.

Apparatus

The laptops were equipped with RED 250 mobile eye trackers (250 Hz) from SensoMotoric Instruments (SMI^{TM}). The displays (15.6 inch and resolution of 1920 x 1080 pixels) were set to the highest brightness level. In combination with a constant testing environment (room illuminance in the experimental room measured by a radiological light sensor, Gossen Mavomax^{TM} illuminance sensor), we achieved an illumination condition of 30 to 40 lux on all displays. The default settings of the SMI Software BeGaze were used to classify the gaze measures (velocity-based algorithm: peak velocity 40°/sec, min. fixation duration 50ms).

7.3.3 Measures

Diagnostic performance

The diagnostic performance was measured by evaluating students' markings, which the students drew on the OPTs using the laptop's mouse to control a digital pen. To assess students' diagnostic performance in the pre- and the post-test, students first saw an OPT for 90 sec. Then, they were asked to mark those regions where either treatments or further follow-up diagnostic

procedures would be warranted. The markings of the students (i.e., circles drawn around suspicious regions) were saved for each OPT and were rated by two trained raters. The raters evaluated students' markings relative to a solution template developed by two experts. The interrater reliabilities for the trained raters compared to an experienced rater were calculated for 20% of the OPTs. The agreement for each of the trained raters compared to the experienced rater for set A (for detection rate: Krippendorff's alphas = 0.97; 0.98, for number of false positives: Krippendorff's alphas = 0.98; 0.94) and set B (for detection rate: Krippendorff's alphas = 0.91; 0.89, for number of false positives: Krippendorff's alphas = 0.96; 0.95) was high, so that the two trained raters continued to code the markings independently. We used the detection rate (percentage of correctly detected anomalies) and the number of false positive markings for the analysis in the pre- and the post-test. The detection rate and the number of false positives were subdivided into two different categories – central and peripheral – depending on their location in the OPT.

Gaze measures

Areas of interest (AOIs) were defined for gaze behavior analysis. The anomaly-AOIs represent the anomalies in the OPTs. The anomaly-AOIs corresponded to the anomalies as they were marked in the solution template by the experts and could be further categorized as located in either peripheral or central areas. If very small anomalies located next to each other represented the same problem (e.g., cavities affecting multiple teeth), they were merged into one larger anomaly-AOI. We used the following gaze measures to analyze the eye-tracking data: the number of fixations on AOIs, fixation time in milliseconds on AOIs, time to first fixation in milliseconds on AOIs, and the overall coverage rate of the OPTs. The latter was determined by dividing each OPT into a grid that consisted of even-sized, rectangular AOIs. For smaller OPTs, we used 14 x 11 rectangular AOIs to build the grid and for bigger OPTs, we used 15 x 11 rectangular AOIs. The area of a single rectangular AOI was 6695 pixels. The coverage rate was determined as the percentage of AOIs fixated within an OPT's grid.

Conceptual knowledge

The two dental medicine experts in the project developed a screening questionnaire to examine students' baseline level of clinical knowledge in dental medicine. The majority of the items came from the Dental School's test item repository and are used in the students' assessments. Newly developed or modified items were reviewed by colleagues from the dental department to further ensure the items' correctness and appropriateness. The questions were

presented on the laptops with the web-based survey software tool Qualtrics. For the 20 multiple-choice questions there were four alternatives and one correct option (e.g., 'Which answer is correct? An apical periodontitis …' answer: '…points towards an endodontic problem.'). There was always one option of 'I cannot answer the question yet / I do not know'. Students got one point for every correct answer and zero points for incorrect answers. The maximum total score was 20 points. Performance was converted into percentage correct.

7.3.4 Procedure

The data collection took place in the Tübingen Digital Teaching Lab at the Leibniz-Institut für Wissensmedien between July 2017 and May 2018. In the intervention group, the pre-tests were conducted approx. three months before the training intervention; the post-test followed immediately after the training intervention. The delays between pre- and post-test in the control group and intervention group were the same. Data collection took place in parallel sessions with up to 30 participants, who worked individually and silently on their assignments. Ethics approval for the study was obtained from the institute's local ethics committee (LEK 2017/016).

At the beginning of all test sessions, the students received written information on the procedure of the experiment and signed a consent form. For the diagnostic task, the students were instructed to seat themselves comfortably in front of the eye-tracker and to not move their head during the task. Then, the students were calibrated with a 13-point calibration before they received the instruction for the diagnostic task. They passed through a short drawing tutorial explaining how to mark anomalies in the OPTs using the drawing plugin tool for Mozilla Firefox™ Browser. Afterwards, the students were informed that they would see the OPTs twice, once in a search and once in a marking phase and were instructed to mark those regions that would require either treatment or follow-up diagnostic procedures. The students also saw instructions regarding cases they should not mark (missing teeth, sufficient treatments, generalized horizontal bone loss, and technical artifacts). Before students entered the search phase, they were shown a fixation cross for 2 seconds. In the search phase, the students were asked to look at the OPTs and search for anomalies. Each OPT was presented for 90 seconds. The search phase was followed by a short instruction reminding the students what anomalies to mark. In the marking phase, the students were asked to mark the detected anomalies with the drawing tool in the OPT. The procedure (instruction - fixation cross – searching phase – instruction – marking phase) was repeated for every OPT.

In the pre-test, the students performed the diagnostic task (10 OPTs of set A for the control group and 20 OPTs of set A and B for the intervention group) followed by the conceptual knowledge test.

Before the post-test, the intervention group received the compare-and-contrast modeling intervention. The students were told that they would see heat map visualizations of their gaze behavior and that of another peer student, where the intensity and location of eye movements were marked in orange. Additionally, the students in the intervention group were informed that a full coverage of OPTs is important and were instructed to compare the peer-model's heat map in the upper part of the screen to their own individual heat map in the lower part. The verbatim instructions were as follows: 'Please use the heat maps to compare your gaze behavior on the OPTs with the gaze behavior of a student in a higher semester. Try to identify similarities and differences in gaze behavior. […] On the next page, you can see the gaze behavior of the student in a higher semester (above) and your own gaze behavior on picture 1 (below). Please look at and compare the two heat maps. You can take as much time as you need. If you then click on 'continue', you will see image 1 (*the OPT*) without heat maps in full size, so that you can view it again. This process will be continued for four more images (*OPTs*). You do not have to mark any conspicuities at this point.' Students were asked to perform the compare-and-contrast task for a total of five heat map combinations with OPTs of set A2. After the training intervention, students were asked whether they had seen differences between their own and the peer-model's heat map. If they had seen any differences, they were asked to briefly describe these differences.

In the post-test, the students performed the diagnostic task (15 OPTs of set A1 and set B for the intervention group and 20 OPTs of set A and B for the control group) followed by the conceptual knowledge test. The students in both groups were recalibrated after five OPTs and could take a short break after 10 OPTs if they wanted.

7.3.5 Data analysis

Missing data

One student in the intervention group had missing values in the conceptual knowledge test due to technical problems. We replaced the missing value of this student by the average group value for semester and time (pre- vs. post-test). All data points available were considered for replacement.

In addition, due to technical problems, the diagnostic performance data of single OPTs were not available for two participants of the control group. Again, we used the remaining data to estimate diagnostic performance values that were used to replace missing values.

Exclusion criteria

For the analysis of the gaze measures, we excluded the first fixation, which is usually residual behavior from the prior fixation cross stimuli before each OPT. Moreover, we excluded the eye tracking data of OPTs with a tracking ratio below 80%. Eye-tracking data of students who reached a tracking ratio above 80% only in half of the OPTs in the pre- or the post-test were excluded from the pre- and the post-test (N = 6 in control group, N = 6 in intervention group). Therefore, there were data from 32 participants in the intervention group and 17 participants in the control group left for analyses of gaze measures.

Analyses

We used linear mixed models to examine the gaze behavior (Hypotheses 1, 2a, 2b & 2c) and generalized linear mixed models for the diagnostic performance (Hypotheses 3). The R package lme4 (Bates et al., 2015) was used for the analysis. The models consisted of the same basic model structure:

$$y_{ijkl} = \beta_0 + \beta_1 \text{Time}_{ij} + \beta_2 \text{Group}_{ik} + \beta_3 \text{Location}_{il} + \beta_4 (\text{Time x Group})_{ijk} + \beta_5 (\text{Time x Location})_{ijl} + \beta_6 (\text{Group x Location})_{ikl} + \beta_7 (\text{Time x Group x Location})_{ijkl} + \beta_8 \text{Conceptual Knowledge}_i + v_{0i} + v_{1i} \text{Time}_{ij} + \varepsilon_{ijkl}$$

y_{ijkl} represents the gaze measure/diagnostic performance of student i. β_0 specifies the intercept across students for the reference categories. The effect of time β_1 (pre-/post-test), the effect of group β_2 (control/intervention), and the effect of location β_3 (central/peripheral anomalies) were included to test the main effects. β_4, β_5, and β_6 each represent two-way interactions between time, group and location; β_7 specifies the three-way interaction between time, group, and location. The effect of the intervention on peripheral and central anomalies was tested by the three-way interaction. With β_8, conceptual knowledge is included as a covariate; v_{0i} specifies the individual intercept for each student and v_{1i} the individual slope over time for each student, see also Appendix A for the measures. Adjustments were made in cases where additional factors had to be included. We used d as an effect size, with d = .20 to .40, d = .50 to .70, and d > .80 corresponding to small, medium and large effects, respectively (Cohen, 1988).

Data transformation

Data distributions of gaze measures were checked by graphical methods (quantile-quantile plots and scatter plots for residuals and predicted values). We used log-transformed values for fixation time and number of fixation (Hypotheses 2a & 2b) because the scatter plots

for residuals and predicted values showed a better distribution for log-transformed values than original values (see Appendix B). For all other measures and analyses, we used the original values due to better distribution in the scatter plot.

7. 4 Results

7.4.1 Comparison between the control and intervention group

Visual coverage of the OPTs (Hypothesis 1)

To test whether the compare-and-contrast intervention would affect the coverage of the OPTs, we augmented the aforementioned basic model with the random factor OPT to account for differences between the OPTs. Moreover, we excluded the factor specifying the location of anomalies and its interactions from the analysis, since the analyses referred to the OPT as a whole rather than to the anomalies contained within them (see Appendix C).

In line with our hypothesis, the results indicated a significant interaction between time and group, Estimate = 4.08, $t(52) = 2.50$, $p = .02$ ($d = .36$). The coverage rate for the intervention group increased slightly from pre- to post-test; however, the coverage rate decreased for the control group (see Table 2). Moreover, students' conceptual knowledge affected their coverage in that better conceptual knowledge was related to a higher coverage rate, Estimate = .45, $t(87) = 2.53$, $p = .01$ ($d = .36$).

Table 2. Means and standard deviation of gaze coverage rates

		Control group		Intervention group	
		Pre-test	Post-test	Pre-test	Post-test
Gaze coverage rate (%)	Mean	49.03	47.09	47.62	48.61
	SD	6.65	6.38	6.31	7.42

Gaze behavior regarding anomalies (Hypotheses 2a, 2b, 2c)

Number of fixations (Hypothesis 2a) The analysis for number of fixations revealed a three-way interaction between time, group and location, Estimate = .33, $t(94) = 3.01$, $p = .003$ ($d = .43$), that was, however, not in the expected direction (see Appendix C). Contrary to our assumption, the number of fixations did not increase for peripheral anomalies in the intervention group. Rather, for both central and peripheral anomalies, the number of fixations decreased in

the intervention group, whereas it increased in the control group (see Figure 5/Table 3). This effect was stronger for peripheral anomalies than for central anomalies.

Fixation time (Hypothesis 2b) We did not find the expected three-way interaction for increase in fixation time on peripheral anomalies due to the intervention. Nevertheless, separate effects for location and an interaction between time and group were found: Figure 5/Table 3 show that students fixated on peripheral anomalies longer than central anomalies, Estimate = 1.09, $t(94) = 13.02$, $p < .001$ ($d = 1.95$). Contrary to our hypothesis, the interaction effect between time and group, Estimate = -.31, $t(113) = -2.66$, $p = .009$ ($d = .37$), was in the opposite direction. In the intervention group the fixation time on anomalies slightly decreased, whereas the fixation time on anomalies increased in the control group.

Time to first fixation (Hypothesis 2c) For the time to first fixating on an anomaly, results revealed the expected significant three-way interaction between time, group and location, Estimate = -6915.16, $t(141) = -2.14$, $p = .03$ ($d = .31$). Figure 5/Table 3 show that students in the intervention group fixated on central anomalies in the post-test later than in the pre-test. In contrast, students in the control group fixated on central anomalies in the post-test sooner compared to the pre-test. This pattern tended to reverse for peripheral anomalies in that students fixated on them earlier after the intervention.

Figure 5

Means and standard errors of number of fixation (left panel), fixation time (middle panel) and time to first fixation of anomalies (right panel) for groups (intervention versus control), time (pre-test versus post-test) and location (central versus peripheral anomalies).

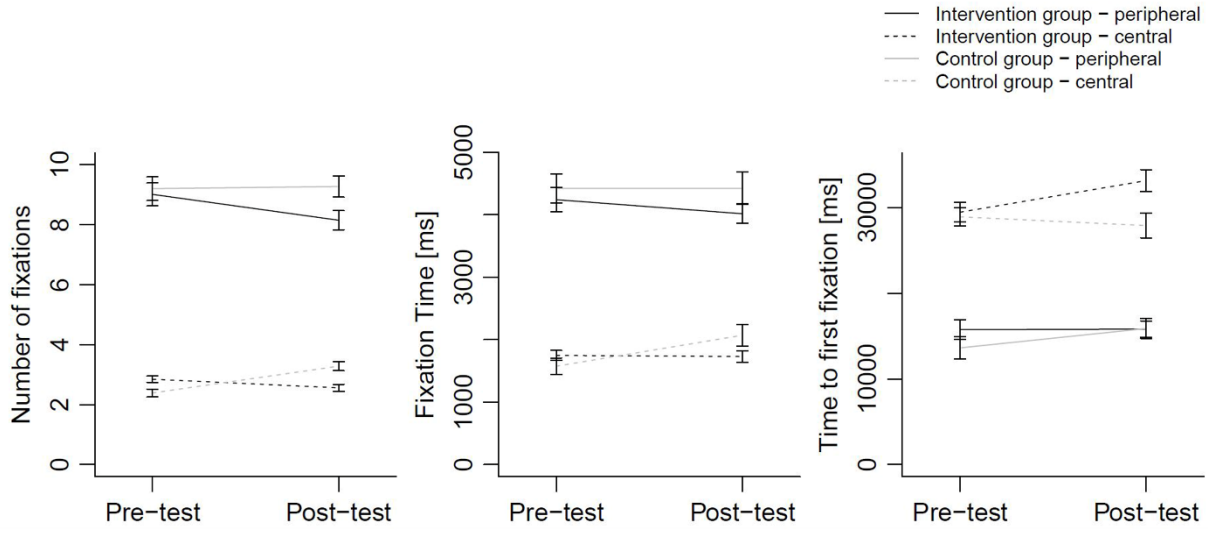


Table 3. Means and standard deviations of gaze measures and diagnostic performance

		Control group				Intervention group			
		Pre-test		Post-test		Pre-test		Post-test	
		central	peripheral	central	peripheral	central	peripheral	central	peripheral
Number of fixations	Mean	2.39	9.21	3.28	9.28	2.85	9.01	2.56	8.15
	SD	0.54	1.62	0.61	1.45	0.63	2.14	0.65	1.80
Fixation time (ms)	Mean	1575.60	4424.18	2070.65	4423.40	1747.96	4241.70	1731.52	4016.95
	SD	539.68	956.70	713.06	1075.59	456.05	1098.39	518.40	884.93
Time to first fixation (ms)	Mean	28927.50	13627.18	27910.00	15889.39	29471.55	15764.38	33156.58	15813.96
	SD	4462.30	5372.14	5967.49	4836.07	6506.46	6501.21	7250.19	5356.15
Detection rate (%)	Mean	51.51	47.06	50.59	51.93	51.69	54.02	54.70	57.02
	SD	8.95	11.63	12.52	14.85	9.85	14.85	10.73	18.79
Number of false positive	Mean	1.12	0.45	1.12	0.55	1.01	0.36	1.54	0.74
markings per OPT	SD	0.58	0.47	0.71	0.45	0.56	0.26	0.89	0.60

Diagnostic performance (Hypothesis 3)

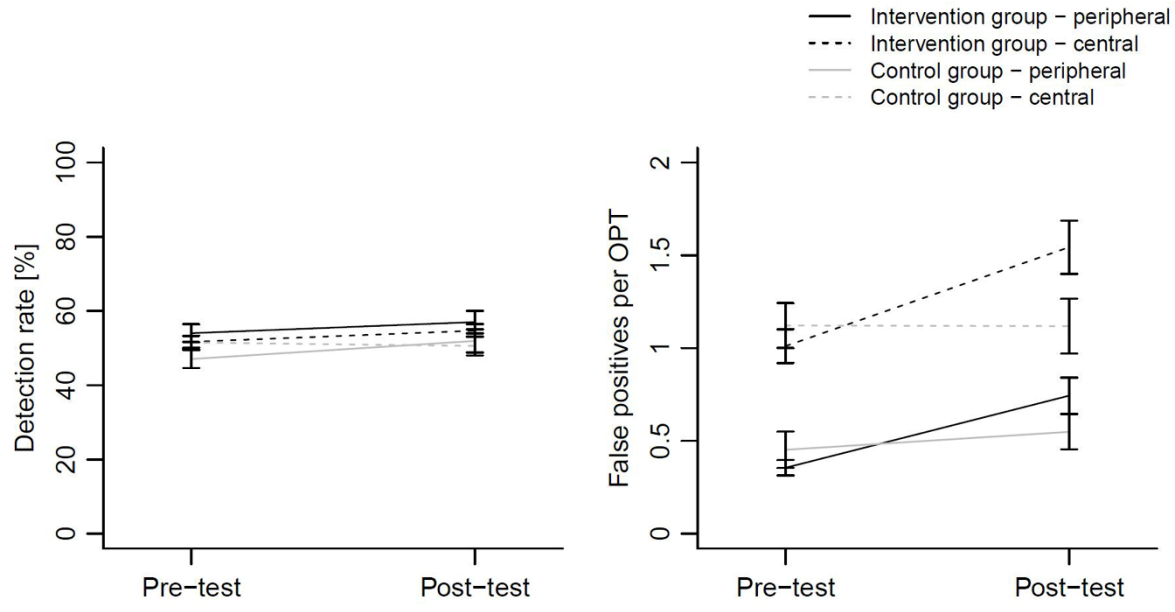
We used a binomial distribution to analyze detection rate and a Poisson distribution to analyze the number of false positive markings. The model was the same as indicated in Appendix A.

Detection rate Against our hypothesis, results did not show an improvement of the detection rate for peripheral anomalies in the intervention group. However, the results revealed a significant interaction between time and group, Odds ratio (OR) = .26, $z = 2.12$, $p = .03$ ($d = .28$). The chance to detect anomalies independent of location slightly increased due to the intervention, but remained stable in the control group (see Appendix D). Figure 6/Table 3 show that this significant interaction was triggered by a slight decrease in the control group for central anomalies.

Number of false positive markings In general, students made more false positive markings in the central area than in the periphery, Estimate = -0.91, $z = -7.88$, $p < .001$ ($d = .97$) (see Appendix D). Contrary to our assumptions, the interaction between time and group showed that the chance to make false positive markings increased in both the intervention and control group, Estimate = .47, $z = 2.95$, $p = .003$ ($d = .38$). However, the increase was stronger in the intervention group (see Figure 6/Table 3).

Figure 6

Means and standard errors of detection rate (left panel) and false positive markings per OPT (right panel) for groups (intervention versus control), time (pre-test versus post-test) and location (central versus peripheral anomalies).



7.4.2 Exploratory analyses

As our intervention did not lead to meaningful improvements of diagnostic performance, we further explored the data to shed light on potential reasons for this finding. One reason could be that our gaze-based intervention only addressed detection errors while recognition and decision-making errors may have occurred as well. To investigate this presumption, we analyzed the distribution of errors resulting from bottom-up processes (detection errors) and errors resulting from top-down processes (recognition and decision-making errors) of all students in the post-test.

Detection errors referred to cases where students neither fixated on nor marked an anomaly. Recognition and decision-making errors were qualified by at least one fixation on an anomaly in combination with a missing marking of that anomaly. The analysis of error types showed that students made on average 0.58 (16%) detection errors and 3.12 (84%) recognition and decision-making errors per OPT. Therefore, students made over five times more errors resulting from top-down than bottom-up processes.

7.5 Discussion

The aim of the study was to improve dental students' search behavior and diagnostic performance in reading OPTs by means of a gaze-based compare-and-contrast modeling intervention. Based on previous research and theories, we hypothesized that students often commit detection errors during OPT interpretation, which could be avoided by supporting them in fuller visual coverage of an OPT. Therefore, students were asked to compare and contrast heat maps of gaze visualizations of a model showing a full gaze coverage and their own heat maps. With this intervention, we aimed at improving students' gaze coverage of OPTs, thereby reducing detection errors.

According to Hypothesis 1, we expected that the gaze coverage of OPTs would increase due to the intervention. Our results support this hypothesis; however, the effect was rather small with a slight increase in the intervention group and a decrease in the control group. In fact, the groups differed regarding coverage of OPTs contained in the post-test only by 1.5%. A reason for this finding may be that our intervention used a very implicit way of increasing gaze coverage. Previous research showed that rather explicit full coverage trainings lead to about 7% difference between control and training group (Kok et al., 2016). Nevertheless, a small increase like the one found in the current study for a short gaze-based intervention with only five heat map comparisons could be also meaningful and a first step to increase coverage when applied

over a longer duration or combined with more explicit instruction as to how to compare the heat maps.

For the search behavior (Hypotheses 2a, 2b, 2c), we assumed that students would fixate on peripheral anomalies more often, for longer, and sooner after the intervention. Most of these predictions were not confirmed; nevertheless, students changed their gaze behavior after they had received the intervention. They fixated on anomalies independent of location less often and for shorter periods of time. A reason for the unexpected behavior could be that students tried to cover OPTs fully, in that they expanded their visual attention, which could be associated with fewer and shorter fixations. The literature on expertise development is ambiguous when it comes to fixation times and number of fixations. Van der Gijp et al. (2017) found that studies equally often report either a decrease or increase of fixation time and number of fixations on relevant areas with higher expertise level. Thus, advanced gaze behavior may be reflected in very different eye tracking result patterns. On the one hand, the fact that students visually process the anomalies less intensely could mean that they were overly occupied with inspecting other areas of the OPTs to obtain full coverage. Therefore, students may not have had enough time to process anomalies sufficiently, leading to a negative diagnostic outcome. On the other hand, it is yet unclear how intensive the visual processing of anomalies must be in order to gain a good diagnostic outcome. Shorter and fewer fixations could mean that students only decided faster. Thus, the intervention lead to a change in the number of fixations and fixation time on anomalies, but further research is needed to specify whether these changes reflect either more efficient or inadequate processing.

The results of the current study confirmed Hypothesis 2c. The intervention lead to later fixations on central anomalies and sooner fixations on peripheral anomalies. Although the effect was small, we can see that the intervention shifts attention towards the peripheral areas of OPTs.

We expected that the intervention would improve diagnostic performance especially, in peripheral areas, as more attention would be directed at these areas (Hypothesis 3). We found that students detected more anomalies independent of location due to the intervention. However, this increase was small, and we are reluctant to interpret this effect as a meaningful improvement. Additionally, an increase in the detection of central anomalies in the control group seems to drive the effect. Thus, we conclude that the intervention did not lead to a meaningful improvement of anomaly detection. Potential reasons for this pattern of results could be traced back to either (a) the only small effects of the intervention on gaze coverage or (b) the type of errors students make:

With the compare-and-contrast intervention, we addressed detection errors, which are errors caused by overlooking anomalies. Our results showed that the intervention lead to only a small increase in OPT coverage, which could be a reason for the very small improvements in detection rate. Another reason might be that – different from what we had expected based on the literature (e.g., Donald and Barnard, 2012) – students do not struggle the most with detection errors, but with recognition and decision-making errors, which were not addressed with our intervention. To investigate this post-hoc explanation, we explored the frequency of detection errors (bottom-up processes) and recognition and decision-making errors (top-down processes) students made during OPT inspection. The results showed that students made about five times more recognition and decision-making errors than detection errors. This large difference could explain why the detection of anomalies did not improve substantially, since our intervention had addressed only a small part of all errors that students made. Thus, future studies in the field of dental radiology should focus more strongly on how to prevent top-down errors (recognition and decision-making), which may be caused by a lack of knowledge about the pathology and the visual characteristics of anomalies. In line with this assumption, research showed that students who learned basic biomedical knowledge improved diagnostic performance of dental radiographs (Baghdady et al., 2009). These observations also reflect the contentious points of theoretical considerations. For decades, two different approaches of problem-based learning in medical education – teaching a problem solving process vs. teaching knowledge - have been discussed (cf. Servant-Miklos, 2019). It is still an open question whether teaching a problem solving process – as we did in this intervention – or teaching knowledge is more beneficial for students (Schmidt and Mamede, 2015). First evaluations suggest that teaching knowledge is more effective (Monteiro et al., 2020; Schmidt and Mamede, 2015). These results also support the view that further studies should focus on improving knowledge in the interpretation of radiographs.

Contrary to our expectations, we found no improvement (decrease) for the marking of false positive errors but an increase caused by the intervention. Students in the intervention group even marked more false positives, whereas students in the control group did not change in committing false positive errors from pre- to post-test. A possible explanation could be that students in the intervention group felt encouraged to find more anomalies. However, due to their potential lack of knowledge regarding characteristic features of anomalies, they did not find more true positive anomalies. Instead, they defined other areas as conspicuous, which resulted in more false positive markings. This phenomenon that interventions lead to more false positive errors is also known in literature (Ganesan et al., 2018). Swensson et al. (1977, 1985)

found that searching for specific anomalies or searching in specific areas lead to an increase in false positive rates. Ganesan et al. (2018) explain this phenomenon by assuming that an intervention may interrupt regular search behavior and therefore lead to more errors.

7.5.1 Limitations

The study has some limitations regarding methods and design. First, five OPTs were used in the pre- and the post-test, which could affect the performance in line with the testing effect (Roedinger III and Karpicke, 2006). However, a potential testing effect should affect the control group in similar ways, but we did not find any improvements for diagnostic performance there. Additionally, previous studies found that observers do not remember the radiographs correctly, suggesting that their repeated use may have little, if any effect on diagnostic performance (Hillard et al., 1985; Myles-Worsley et al., 1988; Ryan et al., 2011).

Second, the rather small sample size in the current study could have contributed to the small effects. However, compared to previous expertise studies that investigated visual search in medical image processing with, on average, only six to eight participants (cf. Gegenfurtner et al., 2011), the sample size in the current study ($N = 61$) is substantial in terms of its statistical power.

Third, the quasi-experimental design with non-randomized groups presents a major limitation of the present study. Due to this design, we cannot exclude that the effects were influenced by cohort differences. However, data from our longitudinal study indicate that the diagnostic performance and most of the gaze measures do not differ between the cohorts. Unfortunately, a randomization of students was not feasible due to different and full schedules of the students. Moreover, the fact that they had to attend the study twice complicated the management and may have decreased students' motivation to participate in the study anyway.

Fourth, the intervention was designed to provide a rather general level of gaze guidance, which may not have been sufficiently specific to improve students' performance. In contrast, dynamic gaze guidance, where attention is directed towards relevant areas on a moment-to-moment basis, has been shown to foster the diagnostic performance of observers (Litchfield et al., 2010; Gegenfurtner et al., 2017). Moreover, dynamic gaze guidance offers information regarding the sequence of inspecting regions and illustrates detailed search strategies. The absence of this information could also have contributed to the pattern of results in the present study. Therefore, it would be worth to investigate in future research, whether dynamic gaze guidance can be helpful for dental students when learning to read OPTs.

7.5.2 Conclusion and implications

In this study, we investigated the effects of a gaze-based intervention on gaze behavior and diagnostic performance in dental students reading OPTs. The intervention changed the gaze behavior of dental students by changing their visual attention but did not improve their diagnostic performance. A potential reason for these findings is that the intervention was developed to address students' detection errors, while post-hoc exploratory analyses showed that students committed more recognition and decision-making errors than detection errors. Thus, interventions focusing only on a full coverage of radiographs appear to not offer the appropriate level of support students would need to improve their diagnostic performance. An alternative training approach would be to focus on teaching visual characteristics of anomalies and basic knowledge of relevant pathology, thereby facilitating top-down processes that help to avoid recognition and decision-making errors (cf. Kok and Jarodzka, 2017b).

8. Study 2:

Comparing radiographs with signaling improves anomaly detection of dental students: An eye-tracking study.

Published online first as: Eder, T. F., Richter, J., Scheiter, K., Huettig, F., & Keutel, C. (in press). Comparing radiographs with signaling improves anomaly detection of dental students: An eye-tracking study. *Applied Cognitive Psychology.* https://dx.doi.org/10.1002/acp.3819

8.1 Abstract

Dental students commit many errors when diagnosing radiographs. To improve performance, students were asked to compare radiographs (with and without disease or with the same disease); relevant structures were highlighted in the radiographs. In a crossover design, students were randomly assigned to two groups differing in training order: Students in the peripheral-central-group (N = 39) were first trained to detect anomalies in the periphery before receiving training on anomalies in the center; the trainings in the central-peripheral-group (N = 39) were reversed. We measured detection rates and gaze behavior before and after each training. The detection rates after the first training revealed differences in line with our expectations; moreover, when accounting for varying difficulty of the tests sets there were within-groups improvements in the peripheral-central group. Unexpectedly, the gaze behavior was unaffected by the intervention. We discuss shorter learning times and sequence effects as potential causes for our findings.

8.2 Introduction

Every patient would be glad if the dentist detected anomalies in dental radiographs as insufficient root fillings or inflammations of the roots and the surrounding bone and prevented her/him from toothache. However, the interpretation of dental radiographs is, as with radiographs from other medical specialties, an error-prone process (Donald & Barnard, 2012; Stheeman et al., 1996). When dentists interpret dental radiographs such as Orthopantomograms (OPTs; dental panoramic radiographs) correctly, this may save a patient's life; for instance, if they detect tumors or calcifications of the carotid artery (Friedlander et al., 2002). Therefore, it is important to conceive effective training methods that steer away dental students from committing errors when interpreting OPTs. So far, unfortunately, evidence-based training methods are still lacking (Kok et al., 2017). Thus, in this study we developed a training to support dental students in interpreting OPTs and evaluated it empirically.

To develop effective training methods, it is important to know first, what dental students need to learn in order to correctly interpret OPTs and, second, where they face the biggest difficulties that need to be targeted by training. Before students are able to interpret radiographs they need to acquire different skills related to perception, analysis, and synthesis processes (van der Gijp et al., 2014). Additionally, specialized knowledge is essential for all three processes – for instance, knowledge of anatomy, of pathology and of epidemiology. With respect to perceiving anomalies, an observer needs to use efficient search strategies, to be able to

discriminate normal from abnormal findings, and to recognize meaningful patterns in a noisy image (van der Gijp et al., 2014). So far, only three studies investigated the perception process of visual search in dental radiographs but without focusing on diagnostic performance of the observers (Grünheid et al., 2013; Hermanson et al., 2018; Turgeon & Lam, 2016). To the best of our knowledge only two studies evaluated training methods to support the perception process and diagnostic performance regarding dental radiographs (Eder, Richter, Scheiter, Keutel, et al., 2021; Richter et al., 2020). The present study addresses this gap in the literature by investigating a training method for dental radiograph interpretation. Verifying the training method's effectiveness in this particular domain is important as dental radiographs and their interpretation differ from radiograph inspection in other medical domains. In particular, dental radiographs typically reveal multiple, different anomalies, which makes their inspection different from, for instance, nodule detection in thorax X-rays, where typically one to two nodules of similar appearance need to be located. In the training described in this paper we aim to improve the perception process and minimize errors at this early stage of acquiring expertise in radiograph interpretation. Kramer et al. (2019) showed that trainings in different visual search tasks share common characteristics. Thus, the training method of this paper might also be applicable to other complex visual search tasks (e.g. baggage screening, lifeguarding) (cf. Kramer et al., 2019).Thereby, the study may have important implications for domains beyond the medical fields.

When it comes to searching for anomalies in radiographs, experts are known to use a global impression of radiographs whereas non-experts, such as dental students, need to resort to a time-consuming search-to-find method (Kundel et al., 2007; Nodine & Mello-Thoms, 2000). Accordingly, many teaching approaches focus on teaching 'systematic viewing' to prevent students from overlooking anomalies (cf. Waite et al., 2019). In this vein, students are taught to cover the full radiograph by looking systematically at every region of a radiograph. However, evaluations of systematic search and full coverage trainings show that while these trainings lead to a more systematic search (van Geel et al., 2017) and/or higher visual coverage rate (Eder, Richter, Scheiter, Keutel, et al., 2021; Kok et al., 2016), students do not improve regarding their diagnostic performance (Eder, Richter, Scheiter, Keutel, et al., 2021; Kok et al., 2016; van Geel et al., 2017). We propose that it is hence necessary to have a closer look at the exact nature of students' errors in order to develop efficient training methods.

Two different kind of errors can occur during radiograph interpretation: False positives and false negatives (Gegenfurtner, Lehtinen, et al., 2017). False positive errors occur when an observer indicates a region as being suspicious that, however, does not refer to an existing

clinical anomaly. False negative errors occur when existing anomalies are not classified as such during the interpretation process. In contrast to false positive errors, false negative errors cannot be corrected in a subsequent treatment process, since in the latter case the patient might not even receive any further clinical diagnosis or treatment as s/he appears healthy. That is the reason why we primarily focus on avoiding false negative errors. The false negative errors during visual search can be further classified into detection, recognition, and decision-making errors (Al-Moteri et al., 2017; Donovan and Litchfield, 2013; Kundel et al., 1978). Detection errors result from bottom-up processes, when an observer did not look at an anomaly. Recognition and decision-making errors, on the other hand, result mainly from top-down processes, when an observer visually attends to an anomaly but does not consider it any further in the diagnostic process. In recognition errors, the observer does not recognize relevant features of the anomaly at all, for example, because of lacking knowledge about pathological/anatomical patterns. In decision-making errors, the observer looks at an anomaly, recognizes the features as suspicious but ultimately decides against their relevance. Originally, the errors types of missed anomalies were classified by gaze behavior: detection errors show no fixation, recognition errors are looked at for less than 1000ms and decision-making errors are looked at for more than 1000ms (Al-Moteri et al., 2017; Donovan & Litchfield, 2013). However, there are also problems regarding the threshold of 1000ms: the threshold seems to be somewhat arbitrary and context-specific (differences for chest radiographs and mammography) (cf. Al-Moteri et al., 2017); moreover, the underlying assumption that increased fixation duration goes along with successful recognition does not seem to hold up (Brunyé et al., 2019). For this reason, we decided to summarize recognition and decision-making errors.

Which of these errors are the most common in radiograph interpretation? Eye-tracking studies evaluated the frequency of these different error types in musculoskeletal radiographs and chest radiographs. Fawver et al. (2020) found that most errors in musculoskeletal radiographs are decision-making errors. The evaluation of chest radiographs showed similar results with some variability, with most errors being decision-making errors followed by recognition errors and detection errors (Donovan and Litchfield, 2013; Kundel et al., 1978; Manning et al., 2004). Hence, while decision errors appear to the most frequent type of errors, there is also quite a bit of variability between studies. However, chest radiographs have different characteristics than OPTs. For instance, the OPTs we used contained up to 26 anomalies in a single OPT, whereas the chest radiographs in the studies mentioned above showed up to five anomalies only. Due to this domain specificity (Gegenfurtner et al., 2011), previous findings cannot be directly transferred to dental radiographs. In our former study, we evaluated what

kind of errors dental students make when diagnosing OPTs (Eder, Richter, Scheiter, Keutel, et al., 2021). We found that the frequency of top-down errors (recognition and decision-making errors) was five times higher than that of bottom-up errors (detection errors). When making false negative errors, the students mostly looked on the anomalies but did not recognize them or made a wrong decision. Thus, to improve students' diagnostic performance it is necessary to decrease the number of top-down errors. Thereby, trainings should support top-down processes by teaching visual characteristics of anomalies and basic knowledge of relevant pathology and anatomy so that students are better able to discriminate anomalies (Eder, Richter, Scheiter, Keutel, et al., 2021; Kok & Jarodzka, 2017b).

One way to foster students' skill in detecting anomalies would be teaching of relevant anatomical and pathological features. Color highlighting can be used to guide students' attention to relevant features. Color highlighting (signaling) is an effective method to guide attention in learning from visual displays (Richter et al., 2016; Scheiter & Eitel, 2015). Furthermore, color highlights in radiology are used in a standard medical teaching book (Pasler & Visser, 2007) or online medical learning tools as AMBOSS (AMBOSS GmbH, 2019) or RadioSurf from University of Bern (Vock & Woermann, 2016). Additionally, verbal descriptions of pathological features are often added in these online medical learning tools. Accordingly, we used color highlights in the OPTs, used as learning material, to guide students' attention to the relevant pathological features and anatomical areas.

Another way to support students in discriminating pathological findings is the compare-and-contrast approach. In a study by Kurtz and Gentner (2013) participants studied depictions of skeletons with and without a comparison standard. The participants detected anomalous features more often when they compared it to a standard depiction. Comparison tasks are also effective for improving diagnostic performance in realistic chest radiographs (Kok, de Bruin, Robben, & van Merriënboer, 2013). Medical students, who compared chest radiographs with a disease against a normal chest radiograph, improved their diagnostic performance for diseases located in a specific area. Furthermore, also comparisons between two radiographs of the same disease or of different diseases were efficient (Kok, de Bruin, Leppink, van Merriënboer, & Robben, 2015). However, overall comparisons showed no benefit compared to a control group. Only when study time was included in the analysis did the comparisons show higher efficiency in the training condition. Thus, the results regarding comparisons for chest radiographs are not as clear-cut. Students may need additional instructional support to benefit from the comparisons. Against this backdrop, in the present study we augmented the compare-and-contrast task with color highlights to foster dental students' ability to detect and discriminate

relevant anomalies during OPT interpretation. Moreover, investigating an enhanced version of a compare-and-contrast training may be particularly relevant for the present domain, since dental radiograph interpretation can be challenging for students given the high number of anomalies.

Eye-tracking can be used to gain insights into visual search processes during OPT interpretation and to evaluate the effects of trainings with compare-and-contrast tasks at a process level. Eye movements consist of two events: fixations, where the gaze is remaining relatively still to allow for intake of the fixated information, and saccades, where the gaze is moving from one focused point to another (cf. Kok & Jarodzka, 2017a). Different measures can be derived from eye movement data. In medical image interpretation the number and time (duration) of fixations on areas of interest (AOIs; i.e., anomalies) as well as the time to first fixate AOIs are commonly used (cf. van der Gijp et al., 2017). The time to first fixation refers to the time it takes a person to look at an AOI such as an anomaly for the first time, indicating the person's ability to quickly detect an anomaly. Longer fixation times and higher frequencies of fixations in a given area typically reflect more intensive processing of that area, for instance, because it is identified as relevant to an upcoming task. We use these measures in the present study to investigate the processes during OPT interpretation and to evaluate the effects of a compare-and-contrast intervention.

8.2.1 The present study

In the present study, we investigate whether an intervention that teaches characteristic features of anomalies could support dental students in diagnosing OPTs. The detection of anomalies was trained in two separate trainings targeted at anomalies that were either located in the central or in the peripheral area of an OPT. The trainings contained written information on the detection of anomalies and compare-and-contrast tasks of OPTs with same-disease comparisons as well as normal vs. disease comparisons. The relevant anatomical and pathological structures in the OPTs were highlighted with colors. The order of the trainings was counterbalanced for participants and their effects were assessed at different time points. For ethical reasons, since the trainings were embedded in the students' regular medical training, we did not use a classic control group but tested both trainings in a crossover design. The participants were part of either the peripheral-central-group (first: training of peripheral anomalies; second: training of central anomalies) or the central-peripheral-group (first: training of central anomalies; second: training of peripheral anomalies). Diagnostic performance was

assessed at three times of measurement (ToM1 – prior to training; ToM2 – directly after first training; ToM3- directly after second training). For the following hypotheses the three-way interactions between location of anomalies (peripheral vs. central), group (central-peripheral- vs. peripheral-central-group) and time (ToM1 vs. ToM2 vs. ToM3) are of relevance.

According to Hypothesis 1, the intervention should support the students in the detection of those types of anomalies that were targeted in the training that they had previously received (in the following these are called peripheral or central anomalies). In particular, we expected that the detection rates for peripheral anomalies should increase in the peripheral-central-group from ToM1 to ToM2 (so after receiving the training regarding peripheral anomalies), whereas the detection rate for the central-peripheral-group should not change (since they did not receive any training regarding peripheral anomalies between the two measurement points). Conversely, from ToM2 to ToM3 we expected an increase in detection rates for peripheral anomalies in the central-peripheral-group, whereas the detection rate should not change for the peripheral-central-group. Analogously, the detection rate for central anomalies should show the reversed pattern: There should be no increase from ToM1 to 2 in the peripheral-central-group, but in the central-peripheral-group; moreover, there should be increases from ToM2 to 3 in the peripheral-central-group, but not in the central-peripheral-group. At ToM3, the detection rates should be the same for both groups regarding peripheral and central anomalies.

Furthermore, the training should yield a faster detection of anomalies and a more intense processing of relevant areas (i.e., areas related to an anomaly), as would be revealed in shorter times to fixate anomalies as well as longer fixation times and higher frequencies of fixations on anomaly-related AOIs, respectively. We expected the intervention to influence the gaze behavior in a way that aligns with the patterns described for detection rate (Hypotheses 2, 3 and 4). In particular, anomalies should be fixated longer (Hypothesis 2, fixation time), more often (Hypothesis 3, number of fixations), and sooner (Hypothesis 4, time to first fixation) when the corresponding type of anomaly (central/peripheral) was trained before compared to types of anomalies that were not trained (peripheral/central), whenever higher detection rates are also expected for them (cf. Hypothesis 1).

The study design, hypotheses and analyses were pre-registered on AsPredicted (https://aspredicted.org/5ei3t.pdf).

8.3 Methods

8.3.1 Participants and design

78 dental students from the University of Tübingen (54 women; mean age = 25.79 years, *SD* = 2.89 years) in their 6th to 10th semester participated voluntarily in the study. The students take a radiology course in their 6th semester and graduate after the 10th semester. The radiology course teaches radiation physics, protection and legislation, imaging techniques, and radiograph interpretation with massed practice of 100 images (half of them were OPTs; cf. Richter et al., 2020). We collected the data of the 6th semester after they had completed the course about radiograph interpretation. Within one cohort, on average 22 students are enrolled – which means that a full survey of 6th to 10th semester students would contain approx. 110 students. Thus, the sample of 78 students covers a representative range of the whole population.

We used a crossover design, where students were randomly assigned to two training groups. The groups received two similar trainings targeting anomalies located in either the central or the peripheral area of OPTs in reversed order; before, between and after they were assessed regarding their diagnostic performance. Students in the central-peripheral-group (N = 39 in semester: 6th n = 9, 7th n = 8, 8th n = 5, 9th n = 5, 10th n = 12, 27 women) passed the training for central anomalies in the first place and the training for peripheral anomalies in the second place. Students in the peripheral-central-group (N = 39 in semester: 6th n = 9, 7th n = 7, 8th n = 5, 9th n = 6, 10th n = 12, 27 women) passed the training in reversed order: first the training for peripheral anomalies and second the training for central anomalies.

8.3.2 Materials

Trainings

The students received two different trainings: one training for types of anomalies in central areas of OPTs (i.e., periapical radiolucency, bone loss, and periodontal gap) and another training for types of anomalies in the periphery, namely, pseudocysts in the maxillary sinuses, osteosclerosis in the jawbones and calcifications of soft tissues in the neck. The trainings contain three steps for each type of anomaly (central training: periapical radiolucency, bone loss, and periodontal gap; peripheral training: pseudocysts in maxillary sinuses, osteosclerosis in the jawbones and calcifications of neck soft tissues).

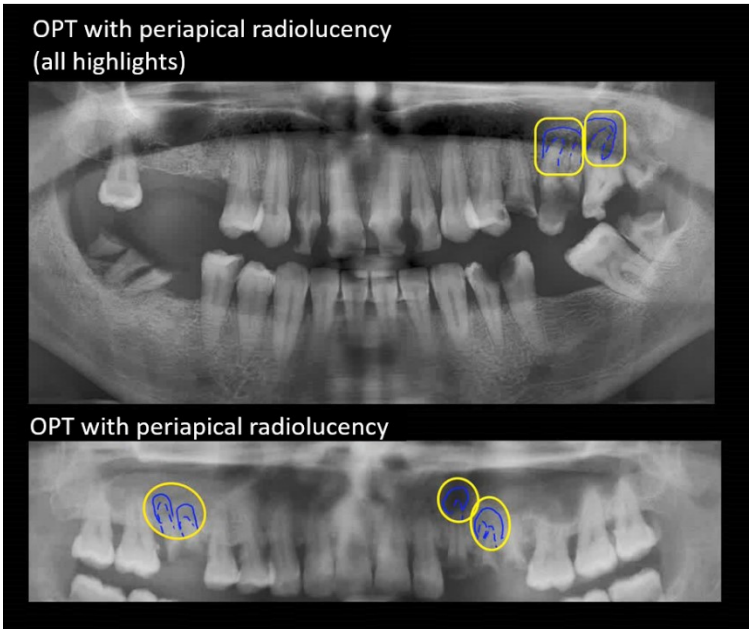
First, a short-written explanation of the anomaly and their visual appearance was given. In the second step, the disease-normal-comparison, the students saw one OPT with anomalies

and one normal OPT (or parts of OPTs) without anomalies and were instructed to compare and contrast the OPTs. The OPT with anomaly was displayed at the top of the screen, the normal OPT at the bottom. In the third step, the same-disease-comparison (Figure 7), students compared two OPTs with the same anomaly (the upper OPT was the same as in the normal-disease comparison).

In the normal-disease and disease-disease comparison students additionally saw colored highlights of those anatomical and pathological structures in OPTs that are necessary to recognize anomalies. The students first saw a pure version of the OPTs without highlights. Second, a version with highlights in blue of the anatomical structure was presented, followed, by, third, another version with highlights of the pathological area, containing the anomaly in yellow. Forth, a version with both anatomical and pathological highlights was shown (Figure 7). After the students passed the four versions by clicking on a "continue" bottom, they could also go back the different versions again by either clicking on bottoms (stating: "no highlights", "highlights anatomy" and "highlights pathology") at the right side of the screen or clicking on the "back" and "continue" bottom at the top of the screen.

Figure 7

Training material. Trainings slide of a same-disease comparison for a version with anatomical (blue) and pathological highlights (yellow).



Apparatus

We used RED 250 mobile eye trackers (250 Hz) from SensoMotoric Instruments (SMI^{TM}) with the default settings of the SMI Software BeGaze to classify the gaze parameters (velocity-

based algorithm: peak velocity 40°/sec, min. fixation duration 50ms). To obtain a constant testing environment with lightning condition of 30 to 40 lux (measured with a Gossen Mavomax^{TM} illuminance sensor), the displays of the laptops (15.6 inch and resolution of 1920 x 1080 pixels) were set to the highest brightness level and the lightning conditions in the experimental room were kept constant.

8.3.3 Measures

Diagnostic performance (detection rate, false positives and top-down errors)

To assess students' diagnostic performance, students were tested at three times of measurements (ToMs) with 5 OPTs each. In total, the tests contained 15 different OPTs from routine checks in the University's Dental Medical hospital with a good picture quality. The OPTs showed a varying number of anomalies; one part of the anomalies were addressed by the trainings (for more details see Table 4). The composition of the OPTs for each time of measurement were selected so that all categories of anomalies addressed by the trainings were represented. Two experts (a maxillofacial radiologist and a prosthodontist both with over 13 years of clinical experience; co-authors of this paper) created solution templates of the OPTs indicating the anomalies. The OPTs were displayed on a laptop with a width between 1,412 and 1,552 pixels and a height of 750 pixels. First, students saw an OPT for 90 sec. Second, they were asked to mark those regions where treatments are needed or regions requiring further clarifications by circling the region with a drawing tool (see procedure for details). Two trained raters rated the OPTs which were saved for all individuals. The raters computed students' markings with a solution template developed by two experts. We used the detection rate (percentage of correctly detected anomalies) for the types of anomalies addressed in the training for the analysis. The number of false positive markings, the probability of marking a fixated anomaly and the ratio of top-down to bottom-up errors were assessed in exploratory analysis. The marking of fixated anomalies describes whether top-down errors (false negatives that were attended with at least one fixation) are reduced in favor of correctly recognizing anomalies (marked and attended with at least one fixation). Therefore, we calculated the relative frequency (recognized anomalies/(recognized anomalies + top-down errors). For the ratio of top-down to bottom-up errors, we calculated the frequency of anomalies not marked but attended with at least one fixation (top-down errors) divided by the frequency of anomalies not marked and not attended with fixations (bottom-up errors) for each ToM.

Gaze parameters

To analyze gaze behavior at the three times of measurements we used areas of interest (AOIs). The AOIs represent anomalies in the OPTs and the AOIs corresponded to the anomalies in the solution template. Very small anomalies located next to each other that represent the same problem were merged into a bigger AOI (see Table 4). We analyzed only the gaze behavior of the search phase (see Procedure) in order to avoid biases due to possible artefacts that may occur by using the marking tool. We used the following gaze parameters to analyze the eye-tracking data: the number of fixations on AOIs, fixation time in milliseconds on fixated AOIs and time to first fixation in milliseconds on AOIs. For all gaze parameters we determined the average per AOI.

Conceptual knowledge

We used conceptual knowledge as a covariate. A screening questionnaire constructed by two dental medicine experts examined students' baseline level of clinical knowledge about dental medicine. The questions were presented on the laptops with the web-based survey software tool Qualtrics. For each of 20 multiple-choice questions, there were four answer alternatives and one correct option (e.g., 'Which answer is correct? An apical periodontitis …' answer: '…is a hint on an endodontic problem.'). One option always stated, 'I cannot answer the question yet / I do not know'. Students got one point for every correct answer and zero points for incorrect answers (maximum total score: 20 points). Performance was converted into percentage correct.

Learning time

The time in minutes that students spent with the training materials was used for exploratory analyses.

Table 4

Characteristics of OPTs at time of measurements

	Number of anomalies							Number of AOIs (gaze parameters)
	Total	Addressed by central training			Addressed by peripheral training			Addressed in the trainings
		Periapical radiolucency	Bone loss	Periodontal gap	Maxillary sinuses	Jawbone	Soft tissues (neck)	
ToM1	56	9	6	2	3	2	2	19
ToM2	57	13	14	2	3	2	5	33
ToM3	39	8	9	2	3	2	2	22

8.3.4 Procedure

We received approval for the study from the local ethics committee (LEK 2017/016). The study took place in the Tübingen Digital Teaching Lab which is a classroom-based laboratory with technology equipment, such as eye trackers, for 30 students at the Leibniz-Institut für Wissensmedien between November 2018 and January 2019. Per session a maximum of 30 students participated. First, students received oral and written information about the procedure of the study and signed a consent form. Then, they were seated individually in front of a laptop with the remote eye tracker attached to it. They were instructed not to move their heads during recording and their eye movements were calibrated with a 13-point calibration.

Before the first ToM, students received a tutorial in which they learned how to draw circles onto the OPTs with a drawing plugin tool for Mozilla Firefox™ Browser to mark regions with anomalies. A written instruction informed the students which anomalies they should mark in the OPTs (i.e., regions where treatments are needed or regions requiring further clarification) and not mark (i.e., missing teeth, sufficient prosthetic and conservative restorations, generalized horizontal bone loss, and technical artefacts). They were also instructed that they would see the OPTs twice in a search and a marking phase directly one after the other. The OPTs and trainings were presented in a browser-based experimental environment developed at Leibniz-Institut für Wissensmedien.

Afterwards, the main part of the study started which contained a test at each time of measurement (ToM) and two trainings; for the sequence see Figure 8.

At ToM1, students' entry diagnostic performance was assessed with 5 OPTs. The students were shown a fixation cross for 2 seconds. In the subsequent search phase, they were asked to search for anomalies in the OPT, which was presented for 90 seconds. After the search phase, a short instruction reminded the students which anomalies to mark (see above). In the subsequent marking phase, the students were asked to mark their detected anomalies in the OPT. This procedure (instruction - fixation cross – search phase – instruction – marking phase) was repeated for the 5 OPTs.

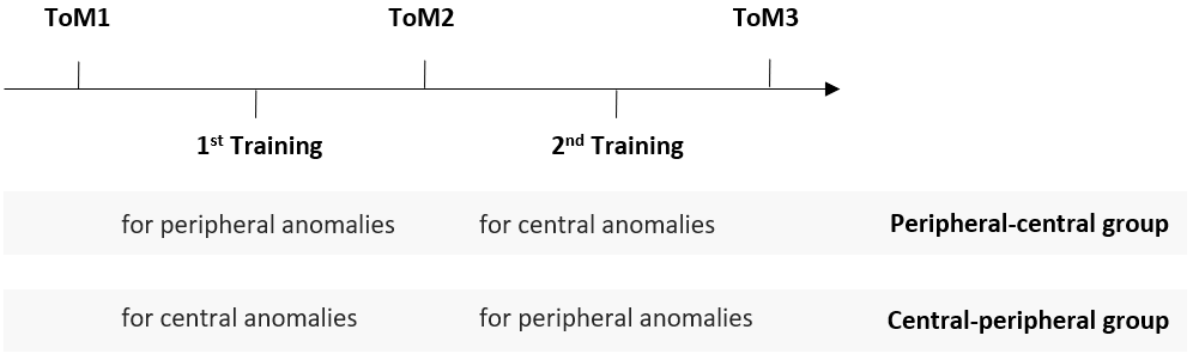
The students received the following instruction for the training before they started the first training: "In the following you will see explanations and visualizations of specific classes of findings (such as caries) in OPTs. First, you will receive descriptions that explain which visual characteristics are used to identify the specific class of findings on an OPT. This is followed by several visualizations that can be helpful for finding anomalies. In the first step you will see a section of an OPT without visualizations. In the second step, anatomical

structures that are relevant for finding anomalies are highlighted in blue. In the third step, the pathological findings are marked with yellow color. Afterwards you will see the anatomical and pathological visualization simultaneously. You also have the option of going back to the different visualizations and taking a closer look at them. You can select the type of visualization on the right side." Then, the students passed the first training (depending on experimental group either for central or peripheral anomalies). The training contained a disease-normal-comparison and a same-disease-comparison for every trained type of anomaly. In the comparisons, students could switch between versions with and without highlights of anatomical structures and / or pathological areas. There were no time constraints for the training and students could go back and forth in the training materials as often as they liked.

After the first training, they had to search and mark anomalies for another set of 5 OPTs (ToM2), going through the same procedures as at ToM1. Then, they received the second training that targeted either peripheral or central anomalies depending on the experimental group. The second training was followed by ToM3. Before ToM2 and 3 the eye tracker was recalibrated. Finally, students completed the conceptual knowledge test. At the end, students were debriefed and received their book vouchers.

Figure 8

Study procedure



8.3.5 Data analysis

Exclusion criteria

For the analysis of the gaze measures, we excluded the first fixation, which is usually residual behavior from the prior fixation cross stimuli before each OPT. Moreover, we excluded the eye tracking data of a single ToM if the calibration of the eye tracker had too low quality (i.e., validation values above .60°: moreover, eye tracking data of single OPTs were discarded

when the tracking ratio was below 80%). If students reached a tracking ratio above 80% only in half of all OPTs, their eye-tracking data was excluded completely (N = 3 in central-peripheral-group, N = 5 in peripheral-central-group). Therefore, 36 participants in the central-peripheral-group and 34 participants in the peripheral-central-group were left for analyses of gaze measures.

Analyses

We used generalized linear mixed models to examine the detection rate and number of fixations (Hypotheses 1 & 3) and linear mixed models for the fixation time and time to first fixation (Hypotheses 2 & 4). We had to deviate from the statement in the preregistration on AsPredicted (see https://aspredicted.org/blind.php?x=a5wz87), because the number of fixations (Hypothesis 3) was not normally distributed; therefore, they were better modelled using generalized linear mixed models than linear mixed models. The R package lme4 (Bates et al., 2015) was used for the analysis. The models consisted of the same basic model structure:

$$y_{ijkln} = \beta_0 + \beta_1 Time_{ij} + \beta_2 Group_{ik} + \beta_3 Location_{il} + \beta_4 (Time \times Group)_{ijk} + \beta_5 (Time \times Location)_{ijl} + \beta_6 (Group \times Location)_{ikl} + \beta_7 (Time \times Group \times Location)_{ijkl} + \beta_8 Conceptual\ Knowledge_i + v_{0i} + v_{1i} Time_{ij} + \eta_{0n} Anomaly + \varepsilon_{ijkln}$$

y_{ijkl} represents the gaze measure/diagnostic performance of student i. β_0 specifies the intercept across students for the reference categories. The effect of time β_1 (ToM1/ToM2/ToM3), the effect of group β_2 (peripheral-central/central-peripheral), and the effect of location β_3 (central/peripheral anomalies) were included to test the main effects. β_4, β_5, and β_6 each represent two-way interactions between time, group and location; β_7 specifies the three-way interaction between time, group, and location. The effect of the intervention on peripheral and central anomalies addressed in the training was tested by the three-way interaction. With β_8, conceptual knowledge is included as a covariate; v_{0i} specifies the individual intercept for each student; v_{1i} specifies the individual slope over time for each student and η_{0n} a random intercept for each anomaly. For calculating the probability of marking a fixated anomaly the covariate and the random intercept for anomaly had to be excluded. Furthermore, we used the estimated-marginal mean function emmeans (Lenth, 2020) for post-hoc pairwise comparisons to specify the results of the (generalized) linear mixed models. P-values of the post-hoc comparison were adjusted with Bonferroni correction.

Data transformation

Data distributions of gaze measures were checked by graphical methods (quantile-quantile plots and scatter plots for residuals and predicted values). We used log-transformed values for fixation time (Hypothesis 2) because the scatter plots for residuals and predicted values and quantile-quantile plots showed a better distribution for log-transformed values than original values. For all other measures and analyses, we used the original values due to better distribution in the scatter plot.

8.4 Results

The means and standard deviations of all measures in the following analyses are displayed in Table 5. The statistical parameters of the analyses are displayed in Table 6.

8.4.1 Detection rate for types of anomalies addressed in the training (Hypothesis 1)

We applied a binomial distribution to analyze the detection rate with a generalized linear mixed model (see above). The results revealed a significant three-way interaction between time, group and location, χ^2 (2) = 69.59, $p < .001$. In line with our assumptions, post-hoc pairwise comparisons showed that the groups differed at ToM2 for peripheral and central anomalies (see Figure 9). The detection rate for peripheral anomalies at ToM2 was higher in the peripheral-central-group, which had received the training for peripheral anomalies before, than in the central-peripheral-group, Estimate = -1.47, z = -5.45, $p < .001$. Analogously, the detection rate for central anomalies was higher in the central-peripheral-group, which had received the training for central anomalies before, than in the peripheral-central-group at ToM2, Estimate = 0.78, z = 3.95, p = .001. That is, as expected students benefitted from "their" training in that they were better able to detect those types of anomalies, for which they had received prior training. Although the pattern in Figure 9 is very similar to the expected hypotheses and shows descriptive increases for peripheral anomalies within the groups after the respective training and a descriptive increase for central anomalies in the peripheral-central-group from ToM2 to ToM3, none of the changes for central and peripheral anomalies between the different time points within the groups were significant in the post-hoc tests, all $p > .05$. Thus, only at ToM2 the intervention affected students' performance, reflected in higher chances to detect peripheral/central anomalies in the group which received the corresponding training for peripheral/central anomalies before. As expected, the groups did not differ at ToM1 or ToM3 regarding the detection of peripheral and central anomalies, all $p > .05$.

Table 5

Means and standard deviations for the diagnostic performance and gaze measures

		central-peripheral-group						peripheral-central-group					
		central			peripheral			central			peripheral		
		ToM1	ToM2	ToM3	ToM1	ToM2	ToM3	ToM1	ToM2	ToM3	ToM1	ToM2	ToM3
Detection rate (%)	M	57.12	56.59	66.26	51.47	63.85	75.46	59.58	44.65	66.67	51.65	82.82	71.06
	SD	15.54	13.65	17.36	17.81	24.35	20.45	17.34	13.52	16.53	20.89	13.17	22.35
Number of fixations	M	1.90	2.77	2.35	5.44	4.87	7.35	2.15	2.79	2.19	5.79	5.40	6.49
	SD	2.12	2.79	2.38	5.28	4.37	7.59	2.21	3.19	2.13	5.69	4.52	7.43
Fixation time (ms)	M	1891	2601	2164	3110	3083	4232	2004	2308	1853	3400	2884	3478
	SD	1894	2231	2264	2748	2187	3797	2045	2224	1708	2808	2612	3343
Fixation time (log	M	7.02	7.44	7.18	7.57	7.69	7.83	7.05	7.24	7.08	7.69	7.52	7.61
transformed)	SD	1.10	1.05	1.08	1.10	0.96	1.16	1.16	1.10	1.01	1.10	1.03	1.16
Time to first fixation	M	31623	29224	32791	24696	19894	21936	32066	29580	30480	22448	20401	20435
(ms)	SD	24009	22714	22068	22466	22325	21944	23195	23571	22579	20921	22670	19726
Z-standardized	M	.12	-.25	-.18	-.19	.09	.29	.25	-.80	-.16	-.18	.96	.06
detection rate	SD	.86	.63	.89	.98	1.12	1.05	.95	.62	.84	1.15	.61	1.15
Probability of marking	M	.62	.57	.63	.56	.70	.65	.58	.46	.66	.60	.81	.60
a fixated anomaly	SD	.17	.14	.18	.18	.17	.21	.20	.15	.16	.18	.14	.22
Ratio of top-down to	M	2.58	4.95	3.29	1.98	1.77	1.75	3.55	4.77	3.26	2.23	1.83	1.50
bottom-up errors	SD	2.17	3.03	2.43	1.23	1.25	0.96	1.99	4.51	1.92	1.12	1.25	0.50

Number of false	M	.94	1.94	3.05	.50	.87	.64	1.14	2.99	3.45	.61	.57	.74
positives per OPT	SD	1.27	2.25	2.53	.98	1.10	.88	1.44	3.06	3.12	1.11	.98	1.12

Figure 9

Detection rate. Means and standard errors of detection rate for location (central versus peripheral anomalies), groups (central-peripheral- versus peripheral-central-group) and time (ToM1, ToM2 versus ToM3).

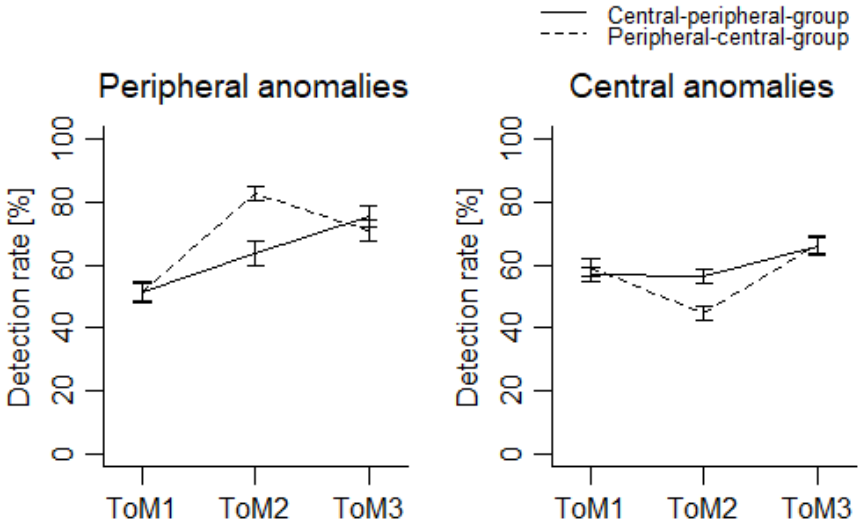


Table 6

Model parameters from the (generalized) linear mixed models of diagnostic performance and gaze measures

	Detection rate	Number of fixations	Fixation Time	Time to first fixation
Fixed effects	χ^2 (df)	χ^2 (df)	χ^2 (df)	χ^2 (df)
Intercept	.55 (1)	1.96 (1)	1095.93 (1) ***	97.12 (1) ***
Time	1.16 (2)	3.16 (2)	4.78 (2)	1.78 (2)
Condition	.54 (1)	1.23 (1)	.05 (1)	.02 (1)
Location	.25 (1)	8.05 (1) **	3.52 (1)	1.98 (1)
Conceptual knowledge	.64 (1)	2.40 (1)	.54 (2)	.00 (1)
Time x Condition	23.84 (2) ***	3.17 (2)	3.78 (2)	1.59 (2)
Time x Location	.71 (2)	14.95 (2) ***	6.02 (2) *	3.08 (2)
Condition x Location	.28 (1)	.33 (1)	.58 (1)	.86 (1)
Time x Condition x Location	69.59 (2) ***	6.28 (2) *	2.40 (2)	.83 (2)
Random effects	*SD*	*SD*	*SD*	*SD*
By subject				
Intercept	.54	.27	.45	5132
Time ToM2 (Correlation intercept)	.58 (-.15)	.20 (-.61)	.39 (-.80)	2470 (-.52)
Time ToM3 (Correlation intercept)	.59 (-.25)	.30 (-.62)	.46 (-.85)	5574 (-.73)
Correlation Time ToM2 and ToM3	.65	.68	.94	.76
By anomaly				
Intercept	1.57	.69	.53	8196
Residual variance (*SD*)	-	-	.90	20557

*p < .05, **p < .01, ***p < .001

8.4.2 Gaze parameters

Fixation time (Hypothesis 2)

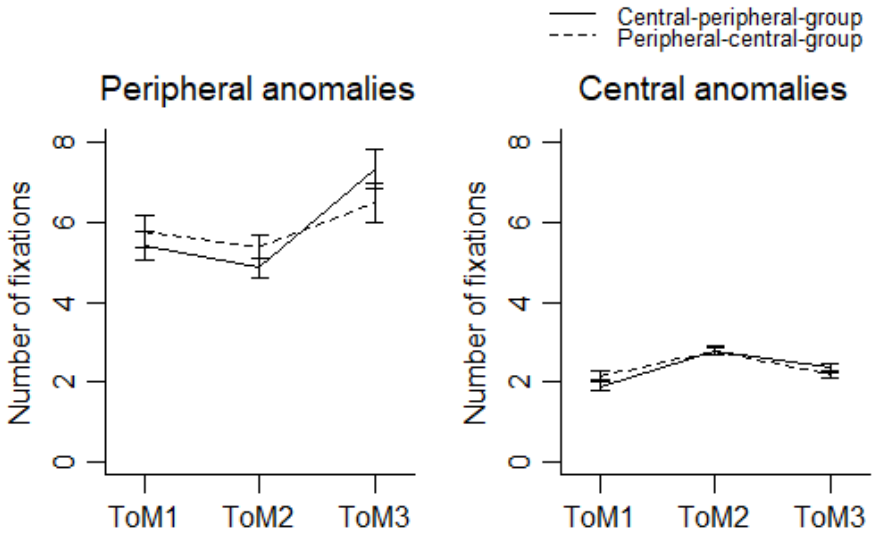
We did not find the expected three-way interaction for fixation time. Nevertheless, we found an interaction between time and location, χ^2 (2) = 6.02, p = .049. The fixation time for central and peripheral anomalies changed differently over time. However, post-hoc pairwise comparison did not indicate significant differences, all $p > .05$, suggesting that the interaction pattern was not sufficiently pronounced to warrant further interpretation.

Number of fixations (Hypothesis 3)

We used a Poisson distribution to analyze the number of fixations with a generalized linear mixed model. The three-way interaction between time, group and location was significant, χ^2 (2) = 6.28, p = .04. Thus, the number of fixations on central and peripheral anomalies changed between the groups over time. The direction of the three-way interaction did not fit to our assumptions. On a descriptive level, Figure 10 shows that the number of fixations only changes slightly for central anomalies, while it increases especially in the central-peripheral group for peripheral anomalies from ToM2 to ToM3. However, post-hoc pairwise comparison did not find significant differences for the changes mentioned descriptively before, all $p > .05$, again suggesting that the interaction was not very pronounced.

Figure 10

Number of fixations on anomalies. Means and standard errors of number of fixations for location (central versus peripheral anomalies), groups (central-peripheral- versus peripheral-central-group) and time (ToM1, ToM2 versus ToM3).



Time to first fixation (Hypothesis 4)

We did neither find a significant three-way interaction, χ^2 (2) = 0.83, p = .66, nor other significant influences regarding the time to first fixation.

8.4.3 Exploratory analysis

Z-standardized detection rate

In the preregistration, only the improvements within the groups were registered as hypotheses, but not the comparisons between the groups. Even though the between-group differences, which we found according to the aforementioned analyses, are a logical consequence of within-groups improvements (in case of equal pretest scores), those could not be revealed. This is most likely due to the differences in absolute difficulty regarding anomaly detection in the OPT sets used at the ToMs. Thus, we ran an additional post-hoc analysis (linear mixed model) using z-standardization of the detection rates at every ToM to eliminate any influence of differences in absolute difficulty (see Table 7). The analysis showed the expected three-way interaction between time, group and location, $\chi 2$ (2) = 25.54, $p < .001$. For peripheral anomalies, the peripheral-central-group showed the expected within-groups improvement in detection rate after the training for peripheral anomalies (comparing ToMs1 and 2 in the peripheral-central-group: Estimate = -1.14, t(228) = -6.20, $p < .001$). Contrary to our expectation, we did not see an increase in detecting peripheral anomalies in the central-peripheral-group from ToM2 to ToM3 (Estimate = -.20, t(226) = -1.09, p = 1.00). As expected, the groups still differ significantly at ToM2 (Estimate = -.86, t(185) = -4.20, $p < .001$), which indicates that the first training for peripheral anomalies was effective. Besides, the detection rate of peripheral anomalies in the peripheral-central-group decreased from ToM2 to ToM3 (Estimate = .90, t(226) = 4.85, $p < .001$), but the descriptive values still show a positive trend when comparing ToM1 before training to ToM3 after the second training for central anomalies.

For central anomalies, the results showed the following pattern: Here, only the within-group comparisons in the peripheral-central-group were significant. They showed a decrease in detection rate from ToM1 to ToM2 (Estimate = 1.05, t(228) = 5.71, $p < .001$); moreover, as expected, there was an increase in detection rate from ToM2 to ToM3 (Estimate = -.63, t(226) = -3.43, p = .01). The between-group comparison at ToM2 was not significant, although the descriptive values showed higher rates in detecting central anomalies after having received the corresponding training (Estimate = .56, t(185) = 2.75, p = .12). Thus, detection rates for central anomalies improved when receiving the second training. However, students appeared to benefit only slightly from the first training, in the sense that they maintained their initial level of

performance after the first training in the central-peripheral-group. Overall, the sequence with first training for peripheral abnormalities and second training for central abnormalities seems to be more effective than vice versa.

Learning Time

We calculated an ANOVA to investigate whether the learning time for the trainings differed between first and second training and whether it was influenced by training sequence. The learning time was longer for the first (M = 3.91 min; SD = 1.30 min) than for the second training (M = 2.55 min, SD = 0.84 min), F (1, 76) = 116.60, p < .001. Additionally, we found a significant interaction between training and group: the decrease in learning time from first to second training was higher for the central-peripheral-group (first training: M = 4.12 min, SD = 1.43 min; second training: M = 2.44 min, SD = 0.76 min) than for the peripheral-central-group (first training: M = 3.70 min, SD = 1.14 min; second training: M = 2.65 min, SD = 0.91 min), F (1, 76) = 6.38, p = .01. These results may indicate why the second training did not lead to significant increases in the detection rate. Enough training time seems to play a role for the effectiveness of the intervention.

Probability of marking a fixated anomaly

Most likely, the intervention would be suited to reduce top-down errors. Thus, we wanted to analyze the detection of anomalies in more detail. Top-down errors (anomalies that were not marked but fixated by the students) should be reduced and transformed into recognized anomalies (marked and fixated) by the intervention. We would expect that the transformation of top-down errors to recognized anomalies, measured in probability of marking a fixated anomaly, is seen in higher probabilities for peripheral/central anomalies after the groups received the corresponding training for peripheral/central anomalies (c.f. Hypothesis 1 for detection rate). We used a generalized linear mixed model to calculate the probability of marking a fixated anomaly. We simplified the model and excluded the conceptual knowledge and the random intercept for anomalies (see Table 7). The results showed the expected three-way interaction between time, group, and location, χ^2 (2) = 17.23, p < .001.

Regarding the peripheral anomalies, post-hoc comparisons showed that the probability to mark fixated anomalies increased for the peripheral-central-group from ToM1 to ToM2 as expected, Estimate = -1.14, z =- 4.80, p < .001. However, from ToM2 to ToM3 this probability decreased while the peripheral-central-group trained the central anomalies, Estimate = 1.09, z = 4.64, p < .001. Contrary to our expectations, the probability to mark fixated peripheral

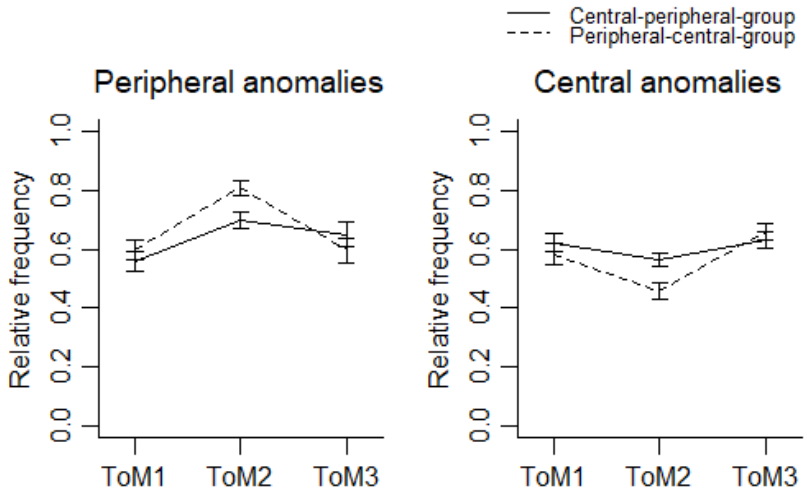
anomalies also increased from ToM1 to ToM2 for the central-peripheral-group, Estimate = -0.69, z = 3.26, p = .02. The increase in the central-peripheral-group was smaller than for the peripheral-central-group (see Figure 11).

For the central anomalies, we observed the expected difference between the groups at ToM2 with higher probabilities in the central-peripheral than the peripheral-central-group, Estimate = 0.51, z = 3.82, p = .002. The change in the probability from ToM1 to ToM2 showed a significant decrease in the peripheral-central-group, Estimate = 0.56, z = 3.91, p = .002, but was not significant for the central-peripheral-group. Thus, in the peripheral-central group, the probability decreased for central anomalies when they received the training for peripheral anomalies before. However, after the peripheral-central-group trained central anomalies in the second training, the probability of marking fixated central anomalies increased as expected from ToM2 to ToM3, Estimate = -0.91, z = -6.98, p < .001.

In total, the intervention seemed to support students in transforming top-down errors to detected anomalies for peripheral anomalies, especially after the first training. The pattern of the results for probability of marking a fixated anomaly (Figure 11) is very similar to the pattern of results for detection rate (Figure 9). This similarity suggests that the reduction in top-down errors likely contributed to the detection rate results.

Figure 11

Probability of marking a fixated anomaly. Means and standard errors of relative frequency (fixated and marked anomalies / (fixated and marked anomalies + fixated and not marked anomalies)) for location (central versus peripheral anomalies), groups (central-peripheral- versus peripheral-central-group) and time (ToM1, ToM2 versus ToM3).



Ratio of top-down to bottom-up errors

To gain deeper insights into the occurrence of the different error types while considering also the different numbers of anomalies per ToM we analyzed the ratio of top-down to bottom-up errors. The descriptive results are shown in Table 5. A post-hoc analysis with linear mixed models did not provide a three-way interaction between time, group, and location, χ^2 (2) = 2.00, p = .37 (see Table 7). Thus, the training did not seem to change the ratio of top-down to bottom-up errors. Furthermore, the results showed that the ratio changed differently over time for peripheral and central anomalies, χ^2 (2) = 8.81, p = .01. While for peripheral anomalies the ratio seems to be stable over time, for central anomalies the ratio was highest at ToM2 in contrast to ToM1 and ToM3. Thus, this increase for central anomalies at ToM2 indicates that the students committed more top-down than bottom-up errors.

Number of false positives

To gain more elaborate insights into the students' diagnostic behavior, we also analyzed the number of false positives (the additional markings of the students that did not meet an anomaly). We calculated a generalized linear mixed model with Poisson distribution to analyze the number of false positive errors. We used a random intercept for the stimuli (OPT) instead of the random intercept for anomalies, which are not predefined for false positives. The results showed a three-way interaction, χ^2 (2) = 26.22, p < .001 (see Table 7). According to Figure 12, the intervention led to more false positive markings in central areas compared to peripheral areas. Especially in the central area, the values differed between the groups at ToM2, Estimate = -0.39, z = -3.08, p = .04, although both groups showed an increase from ToM1 to ToM2, central-peripheral-group: Estimate = -0.95 z = -4.14, p < .001; peripheral-central-group: Estimate = -0.70, z = -3.02, p = .05. The increase was higher for the central-peripheral-group, which received the training for central anomalies, than for the peripheral-central-group. Thus, the trainings did not seem to lead to a more accurate diagnosis in the central area in terms of reducing false positive markings. Students' chance to make false positive errors in the central area also increased significantly between ToM1 and ToM3 for the central-peripheral-group, Estimate = -1.16, z = -5.17, p < .001, and the peripheral-central-group, Estimate = -1.22, z = -5.40, p < .001. Thus, the increase in chance to make false positive errors perseverated even after the second training.

Table 7

Model parameters from the (generalized) linear mixed models of exploratory analyses

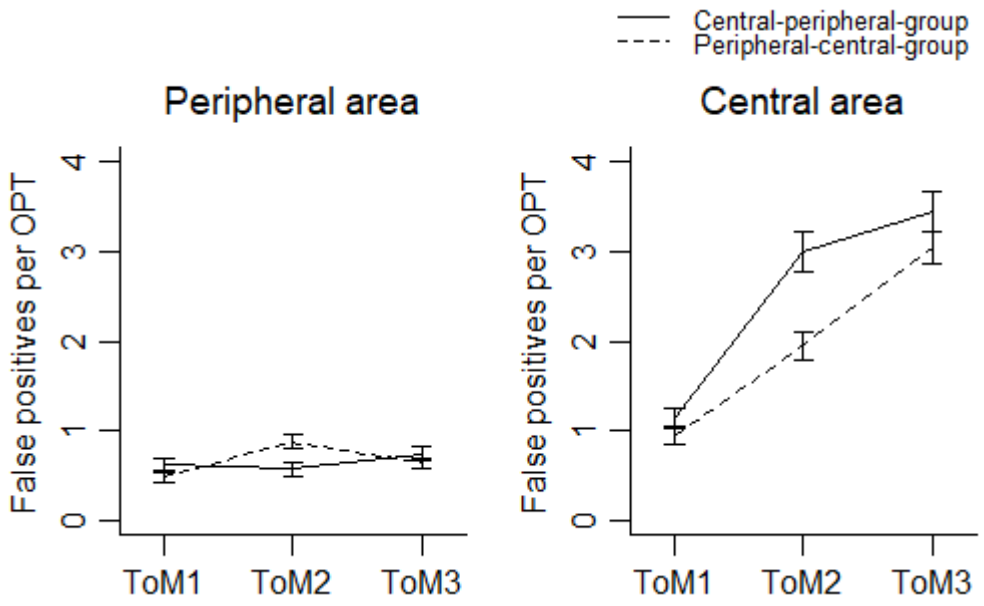
	z-standardized detection rate	Probability of marking a fixated anomaly	Ratio of top-down to bottom-up errors	Number of false positives
Fixed effects	χ^2 (df)	χ^2 (df)	χ^2 (df)	χ^2 (df)
Intercept	.23 (1)	21.71 (2) ***	54.07 (1) ***	1.13 (1)
Time	4.53 (2)	5.17 (2)	12.49 (2) **	29.13 (2) ***
Condition	.35 (1)	.93 (1)	3.15 (1)	.91 (2)
Location	2.84 (1)	2.05 (1)	1.10 (1)	26.40 (1) ***
Conceptual knowledge	1.51 (1)	-	-	8.11 (1) **
Time x Condition	7.92 (2) *	13.77 (2) **	2.21 (2)	9.42 (2) **
Time x Location	10.28 (2) ***	13.66 (2) **	8.81 (2) *	46.53 (2) ***
Condition x Location	.23 (1)	1.44 (1)	.68 (1)	.00 (1)
Time x Condition x Location	25.54 (2) ***	17.23 (2) ***	2.00 (2)	26.22 (2) ***
Random effects	*SD*	*SD*	*SD*	*SD*
By subject				
Intercept	.41	.00	.26	.47
Time ToM2 (Correlation intercept)	.02 (-1.00)	.31 (-)	2.46 (1.00)	.39 (-.43)
Time ToM3 (Correlation intercept)	.10 (1.00)	.36 (-)	.58 (-1.00)	.31 (-.47)
Correlation Time ToM2 and ToM3	-.99	1.00	-1.00	.84

By OPT				
Intercept	-	-	-	.32
Residual variance (*SD*)	.81	-	1.97	-

*p < .05, **p < .01, ***p < .001

Figure 12

False positive errors. Means and standard errors of false positive per OPT for location (central versus peripheral anomalies), groups (central-peripheral- versus peripheral-central-group) and time (ToM1, ToM2 versus ToM3).



8.5 Discussion

This intervention aimed to improve dental students' diagnostic performance and visual search behavior when diagnosing panoramic radiographs (OPTs). To this end, the intervention addressed top-down errors by combining comparing and contrasting OPTs in normal-disease comparisons and disease-disease comparisons with highlighting relevant anatomical and pathological structures. The effect of the intervention was tested in a crossover design with one training for peripheral anomalies and one training for central anomalies, whose order was reversed in the two training groups.

We expected that the intervention led to higher detection rates for these types of anomalies which were trained before (Hypothesis 1). After the first training for peripheral/central anomalies, students showed higher detection rates for peripheral/central anomalies compared to the group which was trained in central/peripheral anomalies. Especially for peripheral anomalies, the detection rate in the group that had trained these types of anomalies was 20% higher than in the group that had trained central anomalies. For central anomalies the difference between the groups was not that high (10%). One possible reason why students did not benefit that much from the training on central anomalies could be their overconfidence in diagnosing central anomalies. Anomalies located in the oral cavity (displayed in the central region of OPTs) are in the natural interest of dentists; Additionally, there is a higher prevalence of anomalies in the oral cavity than in the maxillary sinuses or the neck (displayed in the peripheral region of the OPT) (see Table 4; Constantine et al., 2018; Vallo et al., 2010). Thus, students might overestimate their performance of recognizing central anomalies. This phenomenon of overconfidence as a reason for diagnostic errors is widely known in medicine (Berner & Graber, 2008). Due to the overconfidence in diagnosing central anomalies, students might not have paid sufficient attention on the training material or did not processed it deep enough to benefit from it. The results of the exploratory analysis for false positive errors are also in line with the prior explanation. In general, students had a higher chance to make false positive errors in the oral cavity after they had received the first training. The observation that trainings lead to an increase in false positive errors is also known from the literature (cf. Ganesan, Alakhras, Brennan, & Mello-Thoms, 2018). However, we observed the increase in false positive errors only for central anomalies with especially high increases after the training for central anomalies. Possibly, the increases in false positive errors for central anomalies relates to the overconfidence in detecting central anomalies. Students seem to be very confident in the diagnosis of central anomalies, which may be reinforced by the training.

Thus, they overestimate their diagnostic competence and, triggered by the training, try to find more central anomalies, which lead to more false positive errors.

Regarding the detection rates after the second training and the missing within-group improvements, the descriptive values showed an increase although the results were not significant. Possible reasons for this ineffectiveness, although the trainings were the same as at the first training, relate to learning time (a), the effect size of the intervention (b) or the different levels of difficulty of the anomalies (c). Explorative analyses showed that students took more time to learn in the first training (about 4 min) than in the second training (about 2.5 min) with a higher decrease in the central-peripheral-group than in the peripheral-central-group. Possibly, students did not process the material deep enough to learn from it in the second training, so that the learning time was too short. Another reason refers to the effect size of the intervention. Maybe the effect of the intervention was not strong enough to see significant differences within the groups but only between the groups. The first training yielded differences between the groups but also did not show significant increases within the groups. Differences between the groups were only expected after the first training and not after the second training because of the crossover design. However, also the high adjustment of p-values due to many post-hoc comparisons could lead to non-significant results within the groups. Furthermore, testing our preregistered hypotheses was complicated by the fact that the OPT sets used at the various measurement points differed in the difficulty of detecting anomalies therein. When eliminating the influence of these difficulty differences through z-standardization, the improvements within the groups were also significant in an exploratory analysis. In particular, also the second training showed positive effects for the peripheral-central-group regarding central anomalies. A reason why the second training was not effective for the central-peripheral-group regarding peripheral anomalies might be that this group showed a higher decrease in learning time from first to second training that the peripheral-central group. Another reason could be the order of the trainings. While the trainings improved detection of anomalies in the peripheral-central-group this was not the case for the central-peripheral-group with the reversed order. Thus, the order of training peripheral anomalies first and then central anomalies seems to be more effective than vice versa. Whether this potential influence of training order is meaningful and reliable, needs to be addressed in future studies. The within-group improvements had likely been occluded in the preregistered analyses due to an artefact of the materials. To conclude, the training seems suited to support the detection of anomalies under certain conditions (e.g. enough processing time, training sequence). These results are in line with previous studies

which showed that comparison approaches improved diagnostic performance also in chest radiographs (Kok et al., 2013, 2015).

We also analyzed exploratory if top-down errors, which were targeted with our trainings, are transformed into detected anomalies due to the intervention (frequency of marking fixated anomalies). The pattern of descriptive values was similar to the pattern of results for detection rate, suggesting that the reduction of top-down errors may have contributed to improved detection rate. However, these results should be taken with caution because the analyses did not account for differences in anomaly difficulty and other anomaly characteristics, and the number of bottom-up errors. Further insights into this are offered by changes in the ratio of top-down to bottom-up errors. For peripheral anomalies no changes of this ratio were found. Thus, the training appears to address top-down and bottom-up errors for peripheral anomalies to the same extent. For central anomalies at ToM2, both groups showed a higher ratio than at ToM1 or ToM3, respectively. Thus, students committed more top-down errors relative to bottom-up errors at ToM2. This result indicates that at ToM2 central anomalies seemed to be very difficult to recognize and interpret and is in line with the observation that the training for central anomalies did not lead to an improved detection rate at ToM2. In general, the training is likely to have addressed both top-down and bottom-up errors; however, further studies with a more controlled set of stimuli that are chosen deliberately to investigate the role of different error types and are needed.

Regarding the gaze behavior of the students, we expected that the intervention lead to longer (Hypothesis 2), more (Hypothesis 3) and sooner (Hypothesis 4) fixations on the respective types of anomalies. Against our expectation, we did not find any changes for fixation time (Hypothesis 2) and time to first fixation on anomalies (Hypothesis 4) due to the intervention. We could not find the expected pattern for the number of fixations either (Hypothesis 3). The only change was shown for peripheral anomalies after the second training. Students had a higher chance to make more fixations on peripheral anomalies after they received the second training but no differences between the groups occurred. In general, we also saw higher chances to make fixations on peripheral compared to central anomalies. Thus, students paid more attention on peripheral anomalies. This observation could be related to the detection of anomalies. The intervention was more effective for peripheral than central anomalies. Therefore, the higher attention on peripheral anomalies might helped to support the detection of peripheral anomalies after the training.

One possible explanation why the intervention did not affect gaze behavior as we expected could be that the effect of the trainings was too small to change gaze behavior.

Considering that the trainings only took about four minutes, this is likely too short to substantially change underlying cognitive processes as visual search for anomalies. Another reason could be that the intervention affected only knowledge about visual occurrence of anomalies, pathology, and anatomy but not the visual search itself. Because the intervention was aimed at reducing top-down errors, it did not include, for example, a training of strategic viewing or image coverage, which are more related to the perceptual processes of visual search reflected in eye movements. However, this post-hoc assumption requires further study.

8.5.1 Limitations

The design and methods of this study bring with them some limitations. First, we did not use a classical control group due to ethical and economical aspects. Instead the group which received the training for central/peripheral anomalies served as a point of reference for the results of peripheral/central anomalies trained in the other group. One could argue that the design is not immune against spillover effects. The training that was given to the reference group (e.g. central anomalies) could also have affected the recognition of peripheral anomalies. While these spillover effects would work against our hypothesis, this does not reduce the validity of the results regarding the differences between the groups after the first trainings.

Second, we used different OPTs in the tests at the different times of measurement. While no testing effects can occur through the use of the different OPTs (cf. Roedinger & Karpicke, 2006), it is possible that the OPTs at different times of measurement have different levels of difficulty. We balanced the number of anomalies, which had the same type of anomalies as trained in the trainings, between the OPTs of different times of measurement but the OPTs may still have different characteristics. Although we considered the different characteristics of OPTs by including the anomalies as random effect into the analyses, it cannot be completely excluded that the different OPTs with their different characteristics influenced the results. The characteristics of the OPTs may not only be shown in the anomalies itself but also for example in their composition of anomalies and total number of anomalies. We did not include more characteristics as total number of anomalies into the statistical models because it would have made the models even more complex.

Third, we used separate search and marking phases. While students were instructed to search for anomalies only in the search phase, we cannot rule out the possibility that they continued their search in the marking phase. Thus, gaze behavior computed only from the

search phase might not represent the entire search process and their interpretation is thus limited.

Fourth, we did not distinguish between recognition and decision-making errors. Because a threshold for decision errors in OPTs has not been studied before and is most likely different from the threshold in chest radiographs, we cannot say which type of error was mainly addressed by the training.

Fifth, the training we used in this intervention targeted only the detection of specific anomalies and not a complete diagnosis of OPTs. Therefore, the trainings are not sufficient to teach the diagnosis of anomalies but are one step in the way and could play an important role in teaching visual occurrence of specific anomalies.

Sixth, we used post-hoc pairwise comparisons for a better interpretation of the three-way interactions. The use of post-hoc comparisons brings along the problem of alpha errors. To prevent alpha errors, we adjusted the p-value. This adjustment was strict due to the large number of comparisons within the three-way interactions. This adjustment of p-values could be a reason why we did not find any significant increases within the groups although the descriptive values showed these increases especially for peripheral anomalies.

8.5.2 Conclusion and implication

In this study, we investigated whether comparing radiographs in combination with anatomical and pathological visual highlights improved dental students' detection of anomalies and their gaze behavior. While the students benefitted from the intervention regarding the detection of anomalies, students gaze behavior did not change towards a more targeted or deeper processing of anomalies. The intervention aimed at improving top-down processes by teaching visual occurrence, relevant anatomical and pathological structures of anomalies. A reason for the missing changes of gaze behavior could be that these top-down processes are not reflected in visual search and gaze behavior. To improve detection rates, it seems to be important that students take enough time for the trainings. The training itself is efficient, did not use a lot of technical equipment or personal resources and could be easily implemented in university teaching. For further research it would be interesting to disentangle the effects of the various training components (i.e., radiograph comparisons, anatomical and pathological highlights).

9. Study 3:

I see something you don't: Eye movement modeling examples do not improve anomaly detection in interpreting medical images.

9.1 Abstract

When interpreting medical images such as dental panoramic radiographs (Orthopanthomogram, OPT), errors are frequent. We investigated whether a training with eye movement modeling examples (EMME) and verbal explanations supports dental students in evaluating OPTs. Dental students were randomly assigned to an intervention (N = 42) or a control group (N = 41). The intervention group received the EMME between pre- and post-test. In a laboratory study, we measured students' gaze behavior during evaluating OPTs and the detection rate of anomalies. The training led to fewer, shorter, and later fixations on anomalies and no difference in visual coverage of the OPT. The detection rate of anomalies did not improve. We replicated the latter finding in an online study (N = 31). Students may not have been able to apply the information from the EMME to detect anomalies. The image reading processes changed to more efficient rather than deeper visual search.

9.2 Introduction

When dental students evaluate panoramic radiographs (Orthopanthomogram; OPT), they are known to commit many errors (cf. Eder et al., 2020), indicating deficits in their ability to adequately search and interpret medical images. These deficits still persist in their later working life when they are treating patients as dentists (Stheeman et al., 1996) and may have severe implications for patients. For instance, this may lead to overlooking indications for carotid calcifications which eventually could lead to strokes (Friedlander & Freymiller, 2003). Accordingly, early and effective skill training seems necessary to prevent diagnostic errors in reading dental radiographs. However, well evaluated training methods which could be applied in university teaching are still lacking (Kok et al., 2017). Thus, in the present study, we were interested in evaluating the effectiveness of a training method for medical image processing, which has proven effective for training in other domains that heavily rely on visual information processing, namely, Eye Movement Modeling Examples (EMME).

9.2.1 Diagnostic errors in medical image processing

Two different kind of errors can occur when interpreting medical images such as radiographs (c.f. Gegenfurtner et al., 2017). In the case of false positives, the observer identifies a suspect area as an anomaly even though it is not. On the other hand, false negative errors occur when the observer fails to identify (i.e., overlooks) a true anomaly. We focus on these false negative errors since they are less likely to be corrected later. That is, once an anomaly

has been missed, it will not be considered in the treatment of the patient, unless it causes problems that warrant a second diagnostic procedure. In contrast, false positives errors can be corrected during later diagnostic or treatment procedures.

False negative errors can be further classified concerning the nature of underlying cognitive processing: (a) detection errors, (b) recognition errors, and (c) decision-making errors (Al-Moteri et al., 2017; Donovan & Litchfield, 2013; Kundel et al., 1978). Detection errors (a) occur when an observer does not look at the anomalies and therefore misses them. Recognition errors (b) occur when the observer sees the anomaly but does not recognize the anomaly as such, for example, due to lacking information about the visual features of the anomaly or missing background knowledge. Decision-making errors (c) are similar to recognition errors in that the observer also sees the anomaly, but after deliberation decides against their relevance. The errors result from different cognitive processes: Detection errors rely on bottom-up processes on a perceptual level, whereas recognition and decision-making errors rely on top-down, knowledge-driven processes. Some studies investigated the frequency of the different type of errors that occur during visual search in radiographs with eye-tracking (Donovan & Litchfield, 2013; Eder, Richter, Scheiter, Keutel, et al., 2021; Fawver et al., 2020; Kundel et al., 1978; Manning et al., 2004). Fawver et al. (2020) found mostly decision-making errors in musculoskeletal radiographs, whereas in chest radiographs the frequency for all three types of errors was similar (Donovan & Litchfield, 2013; Kundel et al., 1978; Manning et al., 2004). In OPTs we found that about 80% of the errors result from top-down processes (recognition and decision-making errors) for dental students (Eder, Richter, Scheiter, Keutel, et al., 2021). Thus, the frequency of errors seems to differ between different images and disciplines. In line with this observation, visual search and resulting eye movements depend on many factors such as characteristics of the image, the experimental task, and the specific medical domain (Gegenfurtner et al., 2011; van der Gijp et al., 2017).

9.2.2 Eye-tracking to investigate medical image processing

To investigate these different error types and the visual search of anomalies in radiographs eye-tracking is used (Kok & Jarodzka, 2017a). Eye-tracking allows to assess the visual search behavior of the observers and thereby provides information about how they process the radiograph and anomalies. Many studies have investigated eye movement differences between experts and novices in radiograph interpretation (for an overview see: Gegenfurtner et al., 2011; van der Gijp et al., 2017). Van der Gijp et al. (2017) summarized the important eye movement measures that provide information regarding an observer's expertise level: fixation time on

anomalies (how long the observer looked at an anomaly), fixation count on anomalies (how often the observer looked at an anomaly), time to first fixation on anomalies (when the observer fixate the anomaly for the first time), the number and length of saccades (how frequent and how long were the movements to re-position the eyes) and the coverage of the radiograph (the area of the radiograph fixated by the observer). In general, these expertise studies show that experts typically fixate anomalies sooner and tend to show shorter and fewer fixations than novices (Gegenfurtner et al., 2011; van der Gijp et al., 2017). However, these results cannot be easily applied to the dental student eye movements discussed in this paper, as visual search depends on the domain, task, and level of expertise (Gegenfurtner et al., 2011; van der Gijp et al., 2017). In a study with dental students improvements in diagnostic performance were accompanied by longer fixation durations, more frequent fixations on anomalies, shorter times to first fixation and higher coverage of the image (Richter et al., 2020). Thus, these measures should also be relevant when evaluating trainings for radiograph interpretation in dental students.

9.2.3 Trainings to improve medical image processing

Trainings for visual search tasks, which do not only contain medical images but also, for example, images of airport scanners or beach monitoring as fields of application, can be divided into three approaches: (a) training regarding the use of technology and equipment, (b) training of object identification or (c) training of search strategies (c.f. Kramer et al., 2019). In the following, we focus on a person's ability in detecting anomalies in medical radiographs thereby addressing approaches under the realm of (a) and (b). To this end, we take a closer look at the few evaluated training interventions regarding medical image processing. In doing so, we add the approach of training knowledge to the training of object identification (b), as knowledge about objects, in this case pathology, prevalence and visual features of anomalies, as well as knowledge about the anatomical structure of the image are important to identify anomalies (cf. van der Gijp et al., 2014). In terms of knowledge training, dental students who interpreted dental radiographs improved their diagnostic performance by training basic knowledge such as pathophysiological, anatomical, and physiological information (Baghdady et al., 2009). Trainings of object identification, which aim at teaching visual features and occurrence of anomalies in radiographs, are often designed as compare -and-contrast interventions. In these interventions participants learn to recognize features of anomalies by comparing radiographs with and without pathologies or same/different pathologies (Eder, Richter, Scheiter, Huettig, et al., 2021; Kok et al., 2013, 2015). There is also evidence that combining knowledge trainings

and trainings regarding object identification is more efficient than training both aspects separately (Baghdady et al., 2013). In general, trainings of object identification work well for specific targets (e.g., nodules in the lung) but are less effective when it comes to varied and dissimilar targets (cf. Kramer et al., 2019). As OPTs may reveal a variety of anomalies in different locations, trainings of object identification might not be the best way to improve diagnostic performance regarding the processing of dental images.

Trainings of search strategies (c) aim at improving the process of visually searching for anomalies on radiographs, for instance, by enhancing the area that a person attends to in a radiograph. However, trainings targeting full coverage during visual searching medical images have been shown to not improve diagnostic performance (Eder, Richter, Scheiter, Keutel, et al., 2021; Kok et al., 2016; van Geel et al., 2017). Another learning method supports observers of radiographs in their visual search processes by displaying eye movements of the scanpath in a static image as guidance. In these interventions, the observer either sees his/her own eye movements or that of another person, which are typically represented as a circle or spotlight superimposed onto the image and which correspond with that person's focus of attention. These static gaze interventions have been shown to improve diagnostic performance (Donovan et al., 2008; Kundel et al., 1990; Wedel et al., 2016).

This approach has been developed further by introducing eye movement modeling examples (EMME; cf. Jarodzka et al., 2013; cf. van Gog & Rummel, 2010). EMME consists of videos of another person's eye movements (the model), which are recorded while the model performed the same (learning or problem-solving) task as the observer is supposed to learn. These videos thus show how the model scans the image and fixates areas of interest in order to accomplish the task. By studying EMME, learners are expected to learn from the model how s/he performed the task and to incorporate these processes into their own skill repertoire (cf. observational learning, Bandura, 1971).

EMME have been successfully applied in different contexts: In multimedia learning, EMME supported processing of text and pictures and learning outcome (Mason et al., 2015; 2017; Scheiter et al., 2018). Furthermore, some studies have shown EMME to be effective for enhancing problem-solving and reasoning performance (Jarodzka, van Gog, et al., 2013; Litchfield & Ball, 2011; van Marlen et al., 2018) even though there is also evidence to the contrary (van Gog et al., 2009; van Marlen et al., 2016). In the context of clinical reasoning, EMME supported medical students in learning how to diagnose epileptic seizures in infants based on video recordings of the infants' body movements and led to longer and sooner fixations on relevant areas (Jarodzka et al., 2012). Further studies also investigated the

application of EMME in medical images. Litchfield et al. (2010) showed that EMME of experts and novices improved the search for nodules in chest radiographs especially for novices. EMME also supported diagnostic performance based on interactive medical images such as Computer Tomography (CT) scans or Positron Emission Tomography (PET) scans (for CT and PET: Gegenfurtner et al., 2017; for CT: Seppänen & Gegenfurtner, 2012). Seppänen and Gegenfurtner (2012) used EMME of experts who performed a typical interpreation of a CT Scan. Novice students improved their diagnostic performance and looked more frequent at the relevant areas after studying the EMME video. In the study of Gegenfurtner et al. (2017), novice students only showed a better performance when they interpreted the same case that was shown in the EMME video but not when performing a transfer task. Experts also profited from EMME regarding the transfer task. In addition, the training resulted in relevant areas being looked at more frequently by experts and novices, and the length of fixations generally increased.

In total, the above studies that sucessfully improved visual search showed that EMMEs led to sooner, longer, and more frequent fixations on relevant areas of the stimuli (Gegenfurtner et al., 2017; Jarodzka et al., 2012; Jarodzka, van Gog, et al., 2013; Seppänen & Gegenfurtner, 2012).

There has been a debate among EMME researchers whether or not to provide verbal explanations together with the eye movement videos. Early evidence suggested that adding verbal explanations interferred with learning from EMME, suggesting that at least for knowledge-lean problems verbal and visual guidance may become redundant and thus, may even hinder learning (van Gog et al., 2009). In line with this reasoning, EMME have been shown to be effective even when providing only visual guidance and refraining from additional verbal explanations (Mason et al., 2015; Scheiter et al., 2018). On the other hand, EMME also seem to be enhance task performance when they are accompagnied by verbal explanations (Gegenfurtner, Lehtinen, et al., 2017; Jarodzka et al., 2012; Jarodzka, van Gog, et al., 2013). In the majority of studies, the models were asked to behave didactically when performing the to-be-modelled task and also when verbalizing their reasoning strategies. The didactic explanations can be adapted to the level of learners and make them more understandable for the learner (Isaacs & Clark, 1987). Additionally, the eye movements of experts became more similar to novices when behaving didactically (Emhardt et al., 2020). The authors argue that this behavior may facilitate following the experts task solving behavior for novices. Only in the study by Gegenfurtner et al. (2017) the verbal protocols that reflected the experts' naturally occuring thinking were used as additonal verbal explanations. Regarding medical image

processing, we argue that an expert's explanation that is deliberately tailored towards conveying additional information that aids the interpretation of what is observed (thereby targeting object identification), will be effective to further enhance performance. Accordingly, in the present study we decided to augment an expert's eye movement display with the expert's instructional explanations purposefully designed to facilitate object identification.

9.2.4 The present study

In the present study, we investigated whether the detection of anomalies in OPTs can be improved in dental medical students by means of a targeted training based on EMME. The training comprised eye movement videos of two dental experts (co-authors of the paper) and their didactical verbal explanations that were recorded while they were inspecting OPTs. The students' detection of anomalies and their visual search behavior regarding anomaly detection in OPTs were assessed in a pre- and post-test. In the pre-test, their entry performance was assessed. Then the intervention group received the training between the pre- and the post-test, whereas the control group only performed the post-test without any training. To ensure that the control group had equal access to learning opportunities relevant to their studies though, the control group received the training after the post-test. We expected that the training improves the detection of anomalies, which should be reflected in a higher detection rate of anomalies at the post-test for the intervention group than for the control group (Hypothesis 1). The training should also change the visual search and the gaze behaviour: Students should attend to anomalies more often and longer and detect them sooner after they have received the training. Accordingly, we hypothesized that the training increases the fixation time on anomalies (Hypothesis 2) as well as the number of fixations on anomalies (Hypothesis 3). The time to first fixation on anomalies should decrease due to the training (Hypothesis 4). Furthermore, we expected that the training leads to a more complete visual gaze coverage of OPTs (Hypothesis 5).

Two studies were conducted. Whereas Study 1 aimed at providing a comprehensive test of Hypotheses 1 to 5 in a controlled laboratory set-up, Study 2 was a replication in an online environment conducted during the COVID-19 pandemic. Because it was conducted as a remote study, only the detection of anomalies was assessed. The study design[1], hypotheses and analyses[2] were pre-registered on AsPredicted (https://aspredicted.org/blind.php?x=8ik6re).

[1] Not only 7th to 10th semester students, as originally planned, participated in the study but also 6th semester students to increase the sample size.

[2] For simplicity, we depart from the (generalized) linear mixed model analyses originally planned. Instead we used Covariance analyses, which came to the same or more conservative results than the (generalized) linear mixed models.

9.3 Methods

9.3.1 Participants and design

In Study 1, 83 dental students (58 women, mean age = 25.27 years, *SD* = 2.53), who studied between 6th and 10th semester, of the University of Tübingen participated. According to their curriculum, in the 6th semester, they take a course on radiography where they learn about radiation and imaging techniques as well as how to take and interpret radiographs. The radiology course also entails massed practice where students are required to provide interpretations of 100 dental radiographs. The 6th semester students, who participated in this experiment, had already completed this radiology course. The students graduate after the 10th semester. Until then, there is no further targeted training of medical image interpretation. The students participated voluntarily and gained a 15€ book voucher and received feedback on their diagnostic performance and gaze behavior after the study was completed. The students were randomly assigned to two conditions: the intervention group, which studied EMME videos between the pre- and the post-test, and the control group, which only performed the pre- and the post-test. For ethical reasons, the control group received the EMME videos afterwards. The intervention group consisted of 42 students and the control group consisted of 41 students.

In Study 2, we replicated the first study with 31 dental students (16 women, mean age = 24.27 years, *SD* = 2.91) mostly from the 6th semester (1 student was from the 10th semester) Due to the COVID-19 pandemic, the replication study was conducted as an online study. The participants were randomly assigned to the intervention group or control group. However, four participants did not complete the experiment and had to be excluded, resulting in 13 students in the intervention group and 14 students in the control group.

9.3.2 Material and apparatus

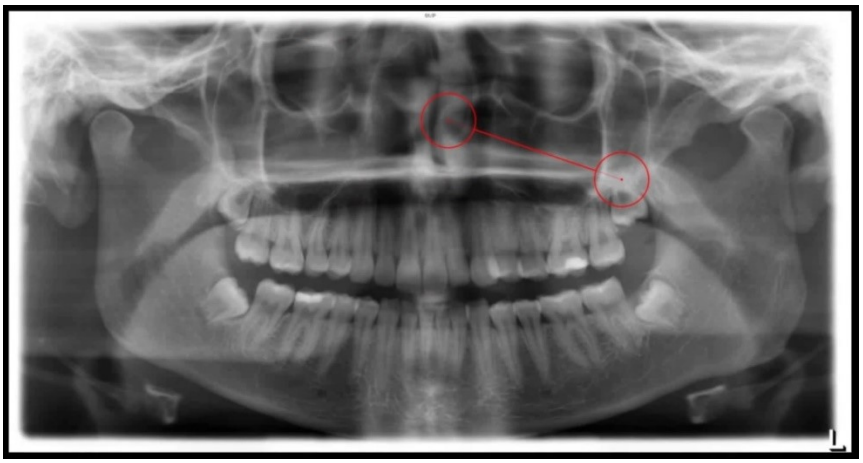
EMME videos

Three EMME videos were used for the training. Two EMME videos (duration: 6 and 5 minutes, respectively) were taken for OPTs with anomalies from a prosthodontist expert (co-authors of this paper with over 13 years of experience). One EMME video (duration: 4 min) was taken for a control OPT without anomalies from a maxillofacial radiologist (co-author of this paper with over 13 years of experience). The EMME videos show the gaze behavior of the experts with red circles representing the fixations and red lines representing the saccades and didactic explanations of the experts (see Figure 13). The cursor (circle) representing the

fixations had a size of 40 px. The eye movements of the experts were recorded with a RED 250 mobile eye tracker (250 Hz) from SensoMotoric Instruments (SMI™) and the SMI Software Experiment Center. We visualized the eye movements with the SMI Software BeGaze and used a velocity-based algorithm to define fixations with peak velocity 40°/s, min. fixation duration 50 ms. The verbal records were didactic explanations which were recorded simultaneously with the eye movements. In the explanations, the experts described how they proceeded as well as what they saw and recognized as suspicious in the OPTs, e.g. “I look at the maxillary sinuses and compare the maxillary sinuses laterally with regard to their opacity and structures. What strikes me is a sharply defined opacity in the right maxillary sinus, which, however, covers all anatomical structures and is therefore more likely to be an artifact or an overlay than an anatomical mass. This is because I do not see this in the patient's left maxillary sinus. The maxillary sinus borders are relatively distinct, and the structures are easily recognizable.” At the end of the EMME videos two still images were shown; one image was the OPT without eye movement superimposition for 10 sec and another image with markings of the anomalies in the OPT for 15 sec (if anomalies were present).

Figure 13

Training material. Screenshot of one EMME video showing fixations (red circles) and saccades (red line connecting the circles).



OPTs

10 OPTs which were recorded during routine checks in the university hospital were used as test stimuli for the diagnostic tasks. They were grouped into two diagnostic tasks with five OPTs in the pre- and the post-test. The OPTs contain between 2 and 26 anomalies. Two experts (co-authors of this paper) examined the OPTs and developed a solution template. All OPTs

showed sufficient clinical image quality. They were displayed with a constant height (750 pixels) and a wide between 1412 and 1552 pixels on a screen of a laptop.

Apparatus

We used 30 RED 250 mobile eye trackers (250 Hz) from SensoMotoric Instruments (SMI™) that were connected to laptops. To classify the gaze measures, we used the default settings of the SMI Software BeGaze (velocity-based algorithm: peak velocity 40°/s, min. fixation duration 50 ms). We enabled a constant illumination condition of 30 – 40 lx on all displays (measured with a radiological light sensor, Grossen Mavomax™ illuminance sensor) by selecting the brightest screen brightness on the laptops. Headphones were used to listen to the EMME.

In the replication online study, students used their own laptops without eye tracking. The students were instructed to use the brightest screen setting.

9.3.3 Measures

Detection of anomalies

We assessed the diagnostic performance as detection rate (percentage of correctly identified anomalies) in the pre- and the post-test. The pre- and the post-test each contained 5 different OPTs from routine checks in the University's Dental Medical hospital with a good image quality. The students saw an OPT for 90 sec. Then, they should mark those regions where treatments are needed or regions requiring further clarifications by drawing circles around suspicious regions using the laptop mouse as input device (see procedure for details). The detection rate was determined by analyzing the markings of anomalies relative to a solution template. Two experts developed the solution template, which indicates the anomalies and their location on the OPTs. Two experienced raters divided the OPTs and matched students' markings with the solution template (for interrater reliability see Eder, Richter, Scheiter, Keutel, et al., 2021).

Gaze measures

We defined areas of interest (AOIs) to analyze the gaze behavior. The AOIs correspond to the anomalies of the solution template (see above). If very small anomalies next to each other corresponded to the same disease (e.g., cavities in several teeth), one big AOI was assigned. The AOI size ranged from 855 px to 22713 px. One OPT contained on average 7.2 AOIs (min:

2 AOIs; max: 14 AOIs). We analyze gaze behavior with respect to total fixation time on AOIs in milliseconds for fixated anomalies, total number of fixations on AOIs, time to first fixations on AOIs and gaze coverage rate of the OPTs. The gaze coverage rate was measured with a grid, which divides the OPT into small, even-sized rectangles. The grid consists of 14 x 11 rectangles for smaller OPTs and of 15 x 11 rectangles for larger OPTs. One rectangle measured 6,695 pixels. The coverage rate was calculated as the percentage of rectangles fixated within an OPT's grid.

We did not collect any gaze measures in the online replication study.

Conceptual knowledge

A screening questionnaire measured students' baseline level of general dental knowledge (e.g., misaligned teeth, root resorptions, soft tissue issues) as a control variable. The two dental medicine experts in the project developed the questionnaire and used mainly questions from the Dental School's test item repository (regularly used for students' assessment). New or modified questions were reviewed by colleagues from the dental apartment to ensure their correctness. The questionnaire contained 20 multiple-choice questions with four alternatives and one correct answer (e.g., 'Which answer is correct? The Stafne cyst is…' answer: '…latent bone cavity in the lower jaw.'). There was always one option stating: 'I cannot answer the question yet/I do not know'. Correct answers were scored with one point, incorrect answers with zero points. The students could reach a maximum score of 20 points, which was transformed into percentage correct.

9.3.4 Procedure

We collected the data in the Tübingen Digital Teaching Lab at the Leibniz-Institut für Wissensmedien between April 2019 and July 2019. Up to 30 students worked individually and silently in parallel sessions.

The students in the intervention group passed the following order of tasks: diagnostic task of the pre-test, studying the EMME, diagnostic task of the post-test and test of conceptual knowledge. In the control group students performed the diagnostic tasks of the pre-and the post-test, then studied the EMME due to ethical reasons, and then worked on the conceptual knowledge test.

In detail, the procedure for students in Study 1 looked as follows: Upon arrival in the lab, the students read information on the procedure of the experiment and signed a consent form. Then, the experimenter instructed the students to seat comfortably in front of the eye-tracker

and to not move their head during recording. The students were calibrated with a 13-point calibration with an automatic validation. They then read the instruction for the diagnostic task (pre-test) and received a short tutorial which explained how to mark anomalies in the OPTs. The students were informed that they should mark regions that would require treatment or follow-up diagnostic procedures. It was also stated that they should not mark specific cases as missing teeth, sufficient prosthetic and conservative restorations, generalized horizontal bone loss, and technical artifacts. Afterwards, students saw each OPT twice: in a search phase and a marking phase. Before the search phase, a fixation cross was presented for 2 sec. In the search phase, students looked at the OPT and searched for anomalies for 90 sec. A short instruction, reminding the students which anomalies to mark followed. In the marking phase, students had unlimited time to mark the detected anomalies in the OPT. This procedure was repeated for every OPT (instruction – fixation cross – search phase – instruction – marking phase). In the pre- and the post-test, the students analyzed five OPTs. The eye tracker was recalibrated before the post-test. The pre- and the post-test were the same for both groups.

After the pre-test, students in the intervention group studied the EMME for which they received the following instruction: “In the following you will see three videos showing the eye movements of an expert recorded during the evaluation of an OPT. You will hear a description of the expert's procedure for reporting and explanations of the pathologies that can be found on the corresponding OPT. At the end of the videos, after the experts interpreted the OPT, you will see the OPTs once again without eye movements. Then the pathological changes (if any) will be highlighted again by means of circles superimposed onto the OPT. The videos will each last about 5 minutes. You cannot interrupt or rewind the videos. Please put on the headphones now to hear the explanations. Click on ‘continue’ to get to the first video.”

Before performing the conceptual knowledge test, students entered their demographic information and we asked them to evaluate learning with the EMME. The vast majority of students described the EMME positively.

In the online replication study, the participants received an invitation via e-mail and were instructed not to perform other tasks while participating in the experiment. The procedure was the same as mentioned above, except that we did not collect eye movements. Thus, instruction regarding eye tracking and calibration were not given.

9.3.5 Data analysis

Missing data and exclusion

Due to technical problems no eye-tracking data of two participants were available in the intervention group.

The first fixation was excluded from analysis of eye tracking data because it is normally residual behavior resulting from the prior fixation cross stimuli. We excluded eye tracking data for participants if deviations in the validation were higher than .6 degree or the data for single OPTs was excluded if tracking ratio fell below 80%. If the tracking ratio fell below 80 % in half or more OPTs or all data was missing in the post-test due to validation values > .6 degree, all eye tracking data were excluded (n = 7 in intervention group, n = 3 in control group). Considering the missing data (see above) and the data exclusion, 33 participants in the intervention group and 38 participants in the control group were included in the analyses of gaze measures.

Analyses

We used covariance analyses to examine the effect of the training at the post-test on detection of anomalies (Hypothesis 1) and on the gaze behavior (Hypotheses 2-5) controlling for the pre-test and conceptual knowledge. The analyses were conducted in R (version 3.5.1) and we used Type II sums of squares. Cohens' d was used as an effect size, with d = .20 to .40, d = .50 to .70, and d > .80 corresponding to small, medium and large effects (Cohen, 1988).

Data transformation

We checked the data distributions of gaze measures with graphical methods (quantile-quantile plots, scatter plots for residuals and predicted values and histograms).

In Study 1, we imputed missing values of five students for gaze behavior on the pre-test with the corresponding means of the students in the same semester. In Study 2, we imputed the missing conceptual knowledge score of one participant with the corresponding mean of all students in Study 2.

9.4 Results

9.4.1 Study 1

Detection of anomalies (Hypothesis 1)

Contrary to our assumptions, we did not find an effect of the training on the detection rate at the post-test after controlling for the detection rate on the pre-test and conceptual knowledge, $F(1, 79) = 1.60$, $p = .21$ ($d = .14$) (for descriptive values see Table 1). The estimated marginal means for the detection rate of the intervention group ($M = 49.30$) and the control group ($M = 52.28$) were very similar. The covariate, detection rate of the pre-test, was significantly related to the detection rate of the post-test, $F(1,79) = 35.82$, $p < .001$ ($d = 1.06$). Conceptual knowledge as covariate did not show a significant influence on the detection rate of the post-test, $F(1,79) = 1.03$, $p = .31$ ($d = -.11$).

Gaze behavior

Fixation time on anomalies (Hypothesis 2). The training affected the fixation time on anomalies at the post-test after controlling for the fixation time on anomalies on the pre-test and the conceptual knowledge, $F(1, 67) = 4.79$, $p = .03$ ($d = .26$) (for descriptive values see Table 1). Contrary to our expectations, the estimated marginal means indicated that the fixation time on anomalies was shorter in the intervention group ($M = 2103.86$ *ms*) than in the control group ($M = 2370.61$ *ms*). The covariate fixation time on anomalies at the pre-test significantly influenced the fixation time on anomalies at the post-test, $F(1, 67) = 24.19$, $p < .001$ ($d = .58$), whereas the covariate conceptual knowledge did not show a significant effect, $F(1, 67) = 0.04$, $p = .85$ ($d = -.02$).

Number of fixations on anomalies (Hypothesis 3). The analysis showed a significant effect of the training on the number of fixations on anomalies at the post-test after controlling for the number of fixations on anomalies on the pre-test and conceptual knowledge, $F(1, 67) = 6.46$, $p = .01$ ($d = .30$) (for descriptive values see Table 1). However, the effect of the training was in the opposite direction as we expected: The estimated marginal means show that the intervention group ($M = 2.79$) fixated fewer on anomalies than the control group ($M = 3.22$). The covariate number of fixations on anomalies at the pre-test showed a significant influence on the number of fixations on anomalies at the post-test, $F(1, 67) = 12.06$, $p < .001$ ($d = .41$), whereas the covariate conceptual knowledge did not have a significant effect, $F(1, 67) = 0.85$, $p = .36$ ($d = .11$).

Time to first fixation on anomalies (Hypothesis 4). The results showed a significant effect of the training on the time to first fixation on anomalies in the post-test when controlling for the

time to first fixation on anomalies in the pre-test and conceptual knowledge, $F(1, 67) = 11.71$, $p = .001$ ($d = -.41$) (for descriptive values see Table 1). Here again, the direction of the effect contrasts with our assumptions. The estimated marginal means for time to first fixate an anomaly in the post-test was later in the intervention group ($M = 32213.91$ *ms*) than in the control group ($M = 27393.76$ *ms*). The covariate time to first fixation on anomalies in the pre-test significantly influenced the time to first fixation on anomalies at the post-test, $F(1, 67) = 10.79$, $p < .002$ ($d = .39$), whereas the covariate conceptual knowledge did not show a significant effect, $F(1, 67) = 0.70$, $p = .40$ ($d = -.10$).

Table 8

Means and standard deviations of observed detection of anomalies and gaze behavior

	Intervention group		Control group	
	Pre-test	Post-test	Pre-test	Post-test
Detection rate (%)				
Mean	66.33	50.29	62.26	51.17
SD	15.71	13.12	16.29	12.81
Fixation time (ms)				
Mean	2598.88	2182.16	2342.26	2359.64
SD	2908.58	2230.18	2452.02	2242.75
Number of fixations				
Mean	3.23	2.79	3.12	3.20
SD	3.68	3.23	3.51	3.57
Time to first fixation (ms)				
Mean	26236.55	32194.22	27121.94	27411.76
SD	22055.58	24212.49	23174.63	22689.71
Gaze coverage (%)				
Mean	47.10	46.87	49.20	46.21
SD	7.19	6.19	6.27	7.16
Study 2 - detection rate (%)				
Mean	35.28	34.14	36.78	28.86
SD	14.55	11.67	17.00	15.25

Gaze coverage (Hypothesis 5). Contrary to our expectations, the training did not affect the gaze coverage at the post-test after controlling for gaze coverage on the pre-test and

conceptual knowledge, $F(1, 67) = 2.14$, $p = .15$ ($d = -.17$) (for descriptive values see Table 1). The estimated marginal means for gaze coverage in the intervention group ($M = 47.21$) and the control group ($M = 45.65$) did not differ much. The covariate, gaze coverage of the pre-test, was significantly related to the gaze coverage of the post-test, $F(1,67) = 40.55$, $p < .001$ ($d = .76$), while the covariate conceptual knowledge was not significant, $F(1,67) = .74$, $p = .39$ ($d = -.10$).

9.4.2 Explorative analysis: types of errors

The training aimed at teaching search strategies and reducing bottom-up errors (anomalies not marked and not fixated by the students) as well as top-down errors (anomalies not marked but fixated by the students at least once). To investigate how the training affected the distribution of errors, we analyzed the frequency of the error types for the pre- and the post-test. The descriptive values show no meaningful changes in the frequency of error types (Table 9).

Table 9

Frequency of error types

	Intervention group		Control group	
	Pre-test	Post-test	Pre-test	Post-test
Frequency of error types (%)				
Bottom-up errors	16.60	18.89	16.14	14.89
Top-down errors	83.40	81.11	83.86	85.11

9.4.3 Study 2 – online replication study

Detection of anomalies (Hypothesis 1)

Against our assumptions, but in line with findings from Study 1, we did not find an effect of the training on the detection rate at the post-test after controlling for the detection rate on the pre-test and conceptual knowledge, $F(1,23) = 2.11$, $p = .16$ ($d = -.28$) (for descriptive values see Table 1). The estimated marginal means for the detection rate of the intervention group ($M = 35.13$) and the control group ($M = 29.57$) were similar. The covariate, detection rate of the pre-test, significantly influenced the detection rate of the post-test, $F(1,23) = 21.52$, $p < .001$ (d

= .89). Conceptual knowledge as covariate did not significantly influence the detection rate of the post-test, $F(1,23) = 1.25$, $p = .28$ ($d = .21$).

9.5 Discussion

In this study, we aimed to evaluate a training that uses EMME and didactic verbal explanation of experts during OPT interpretation. Dental students studied the EMME videos plus explanations. We expected that the training enhances their visual search and improves their detection of anomalies compared to dental students who did not receive the training.

Regarding Hypothesis 1, we assumed that dental students benefit from the training and show an improved detection of anomalies. Our results did not support this hypothesis because the detection rate did not differ between intervention and control group. In accordance with this result, we did not find any effect of the training on detection rate in the online replication study either. The detection rate in the online replication study was substantially lower than in Study 1. Reasons for this might be the level of semester (students in the online replication study were mostly from the lowest (sixth) semester) or the difficulties, which come along with an online study, which does not provide a strict controlled setting.

The null results contradict findings from previous EMME studies in radiograph interpretation (Gegenfurtner, Lehtinen, et al., 2017; Litchfield et al., 2010; Seppänen & Gegenfurtner, 2012). However, these studies used EMME in chest radiographs and, to best to our knowledge, this study is the first to investigate EMME in dental radiographs. The domains differ a lot in the number of anomalies per radiographs. While chest radiographs just show a small number of anomalies, OPTs can show many anomalies (up to 26 in the OPTs used in this study). Thus, domain matters (Gegenfurtner et al., 2011) and the characteristics of OPTs with many anomalies could have led to cognitive overload (see the following section).

We will discuss three possible explanations for the null results in the following: First, the training might show long term effects which we did not measure in this study. Kramer et al. (2019) argue that trainings of search strategies may not show short-term but long-term effects. As EMME model the search strategy of experts and students should benefit from it by adopting the strategies, it is possible that the training will show postponed effects. To benefit from search strategy training, it is necessary for students to have sufficient knowledge of anomaly decision making (Kramer et al., 2019). Because decision-making skills develop slowly, it is possible that search strategy training will have an effect only in the long run. Training that only improves

performance in the long term can also be found in other fields (e.g. woking memory of children: Ramani et al., 2020).

Second, the students saw three EMME videos with explanations, which lasted in total 15 min. This could be a long time for students to pay attention when confronted with new visualizations as eye movements of the experts, new OPTs, and terminologies that all had to be internalized and remembered. Maybe the students got cognitively overloaded due the wealth of information (e.g., high number of anomalies) and could not implement the new strategies and information in their own search for anomalies.

Third, maybe the training does not address the problems students have when evaluating OPTs. The students could miss background information which is necessary to recognize anomalies. As the different error types can provide insight into students' problems, we calculated the frequency of these error types in an exploratory analysis. The training did not affect the frequency of the error types; however, the analysis show that students have more problems with top-down errors than bottom-up errors. This observation fits with our findings in a previous study (Eder, Richter, Scheiter, Keutel, et al., 2021). We had expected the training to address both bottom-up and top-down errors with the EMME and didactic verbal explanations. Although the nature of the training is more a training of search strategies, we had hoped to address also object identification with the didactic verbal explanations. However, the training might not have a strong effect on object identification and the reduction of top-down errors and thereby did not address students' problems in evaluating OPTs (c.f. Kok & Jarodzka, 2017b). When basic knowledge is missing, the training could have overstrained the students with their previous image interpretation skills. In medical education there is a general discussion about whether problem solving trainings as strategy trainings or knowledge trainings have a stronger effect on students' learning (Schmidt & Mamede, 2015). There is evidence that at the general level of medical education knowledge trainings work better (Monteiro et al., 2020; Schmidt & Mamede, 2015). It is possible that these general results also apply to strategy trainings in medical image interpretation.

In relation to the gaze behavior, we expected that the training leads to longer (Hypothesis 2), more (Hypothesis 3) and sooner (Hypothesis 4) fixations on anomalies and a higher visual coverage of the OPTs (Hypothesis 5). However, we found exactly the opposite pattern: shorter, fewer and later fixations on anomalies in the intervention group compared to the control group and no differences in gaze coverage between the groups. We thought that the students would enhance their visual processes as this was the case in the study of Richter et al. (2020). However, the results also contradict findings from previous EMME research in visual search tasks, which

showed longer (Jarodzka et al., 2012; Jarodzka, van Gog, et al., 2013), more (Gegenfurtner, Lehtinen, et al., 2017; Seppänen & Gegenfurtner, 2012) or sooner (Jarodzka et al., 2012; Jarodzka, van Gog, et al., 2013) fixations on relevant areas. We will give two possible explanations on the resulting pattern of the present study: First, the findings can be interpreted as suggesting a more efficient visual search. While students fixated the anomalies shorter and less frequently, their detection rate did not decrease. Thus, anomalies were detected with fewer and shorter fixations. Additionally, the results of the gaze coverage (Hypothesis 5), which did not change to a higher coverage rate, fits with this explanation. A reason for the more efficient visual search could be that students have adopted the gaze behavior of the experts. In general, experts have been shown to reveal shorter and fewer fixations on the radiographs, whereas results for fixations on anomalies are inconsistent (van der Gijp et al., 2017). In contrast, investigations of the first fixations on anomaly in experts are unambiguous and show sooner fixations for experts (van der Gijp et al., 2017). Thus, an adoption of their behavior would not explain why students fixated anomalies later. Second, students might have expanded their visual search and therefore looked less often, shorter, and later at specific areas as the anomalies. The results are in line with gaze behavior we found in another training study (Eder, Richter, Scheiter, Keutel, et al., 2021). There, students also showed shorter and fewer fixations on anomalies after they received a search training.

Overall, it can be said that gaze behavior changed as a result of the training in that students showed more efficient visual search.

9.5.1 Limitations

The study has some limitations in methods and design. First, we used a control group, which did not have a filler task or anything else while the intervention group receive a training. Normally, the control group would be a weak control group. However, even with this weak control group the training did not affect the detection rate of anomalies. Second, we did not use individual groups that received only the eye movements of the EMMEs or the didactic verbal explanation of the models and thus could not distinguish between effects of the individual components of this training. If this training had shown positive effects on anomaly detection, this study would have been an interesting starting point to investigate these individual components. Third, we did not use a follow-up test. Thus, we do not know if long-term effects of the training exist. Further studies are needed to evaluate if the search strategy training with EMME has long-term effects instead of short-term effects (c.f. Kramer et al., 2019). Forth, at

the end of the EMME videos, the anomalies were highlighted by markings. This highlighting adds an additional component to the training that is not normally given in EMME. Our intent was to provide dental students with a summary of the EMME in the form of highlighted anomalies. These highlights could reduce the cognitive load and allow students to reflect on the characteristic features of the anomalies. This reflection could help to build knowledge and reduce top-down errors. However, the highlighting interferes with pure effects of EMME and limits informative value of pure EMME. Therefore, the results of this study must be viewed as effects of a training with different components (EMME, verbal explanation, and highlighting). Fifth, we used different OPTs at the first and second time of measurement. Thus, it might be that the OPTs and anomalies have different levels of difficulty. However, the different OPTs also offer advantages of high ecological validity and absence of testing effects (cf. Roedinger III & Karpicke, 2006). Sixth, the number of participants in the online replication study (Study 2) was small. Therefore, interpretation of the results is limited.

9.5.2 Conclusions and implications

In this study, we investigated if EMME with verbal didactical explanations of experts support dental students in reading OPTs. The training did not improve the detection of anomalies neither in the original study nor in the replication online study. The gaze behavior of the students in the intervention group changed after studying the EMME videos to a more efficient visual search and not, as we expected, to a deeper processing of anomalies.

It is possible that the effects of the training can only be observed in the long term, that the training was too demanding for the dental students or did not sufficiently address the problems faced by students. Thus, further studies are needed to evaluate why the training was not beneficial for the dental students. It would be interesting to measure cognitive load and evaluate the EMME videos qualitatively in more detail. This could give an indication of the degree to which bottom-up and top-down errors are addressed with the training.

10. General discussion

Interpreting panoramic radiographs (OPTs) is an error-prone process for dental students (Richter et al., 2020). These errors in interpreting radiographs persist after graduation from university (Stheeman et al., 1996), which highlights the urgent need for evaluated training methods to improve anomaly detection in dental radiographs. To the best of my knowledge, there are no studies to date that have examined specific training methods for OPT interpretation. Thus, this thesis investigates how dental students can be supported in interpreting panoramic radiographs. This dissertation researched methods that were used as training for other radiographs (e.g. chest radiographs) (EMME in Study 3), combined them in a new way (compare-and-contrast method with individualized gaze feedback in Study 1) and added new components (signaling in radiograph comparisons in Study 2).

Three training methods were evaluated in three different studies and were expected to improve detection of anomalies and lead to more intensive visual processing. Study 1 evaluated an individual gaze-based compare-and-contrast intervention with a control and intervention group. This intervention aimed to expand dental students' visual search (training of search strategies) and thereby reduce detection errors of anomalies located in the periphery. Dental students in the intervention group received comparisons of static heatmaps showing the visual search behavior of a peer model with full coverage search on an OPT and their own visual search behavior of the same OPT. Study 2 focused on the reduction of recognition and decision errors. The comparison of radiographs with and without diseases/with same disease anomalies are used to train the identification of anomalies. The comparisons were supported by colored highlights of relevant anatomical/pathological areas and descriptions of characteristic features of anomalies. This training method was evaluated in a crossover design with one training session for recognizing peripheral and another training session for recognizing central anomalies. The order of the two training sessions differed for the two groups to which the students were assigned. Study 3 aimed to reduce detection, recognition and decision errors with a training method that combined training of visual search strategies and object identification. EMME of experts aimed to teach visual search strategies and promote the identification of anomalies with its simultaneous didactic verbal explanation. Dental students in the intervention group saw the EMME videos between a pre- and post-test while the control group only performed a pre- and post-test.

The following summarizes the results of the three studies for the diagnostic performance and the gaze behavior before discussing how to improve diagnostic performance and visual

search in medical image processing and addressing strengths, limitations, further directions and implications of this dissertation.

10.1 Summary of results

In this section, the results of the studies are structured based on the hypotheses. Thus, the results are summarized for diagnostic performance from all studies (separate section for detection rate (H1a), false positives (H1b), and error types (H1c)) before reviewing the results of gaze behavior (sections for visual coverage (H2), fixation time on anomalies (H3), number of fixations on anomalies (H4), and time to first fixation on anomalies (H5)).

In general, the intervention methods in all three studies should improve the detection of anomalies (H1a). In Study 1, the intervention of gaze feedback comparisons should improve the detection of peripheral anomalies. The results did not show such a specific support of the intervention for peripheral anomalies. Instead, a minor change between the groups over time, irrespective of the location of anomalies was found, suggesting that the intervention slightly supported the detection rate in general (peripheral and central anomalies). However, this effect seemed to be mainly triggered by a decreased chance to detect central anomalies in the control group. Thus, the intervention of Study 1 cannot be considered beneficial for anomaly detection despite the found differences. In Study 2, the detection rate of those anomalies that had been addressed in the first training improved. Reasons why the detection rate only improved after the first but not after the second training session were addressed with further exploratory analyses. The time spent on training might play a role, as the training time on average was longer for the first training session than for the second training session. A more important reason for explaining the differences in detection appears to be the difficulty of the test items, as the OPTs in the second test contained anomalies that were more difficult to recognize. Taking these differences in difficulty into account, post-hoc analyses showed that students benefited from the training when they trained recognition of peripheral anomalies first and then recognition of central anomalies as revealed in detection rates. This was not the case when central anomaly recognition was trained first and peripheral anomaly recognition second. One explanation for this sequence effect could be the prevalence and resulting overconfidence for central anomalies. If recognition of central anomalies, which typically have a higher prevalence than peripheral anomalies, are trained first, students may already be familiar with these anomalies. This familiarity might have led to an overconfidence in students' image interpretation skills and thereby to a poorer processing of the training and weaker performance (Berner & Graber, 2008).

Thus, students also might have had the feeling that they cannot profit from training recognition of central anomalies. These attenuating effects could then carry over to the second training session for recognition of peripheral anomalies. In line with this assumption, the training time decreased more in the group that trained recognition of peripheral anomalies in the second training session. In contrast, students who had trained recognition of peripheral anomalies first and potentially felt they had learned a lot because they were unaware of these low-prevalence anomalies may also have been more motivated in the second training session. In Study 3, students did not benefit from the training with EMME videos regarding the detection of anomalies. This finding was found in a lab study and replicated in an online study.

Regarding the number of false positive errors (H1b), in Study 1 the intervention led to a higher increase in false positive errors than in the control group. The number of false positives were in general higher for central than for peripheral anomalies. In Study 2, the number of false positives also increased for central anomalies after both training sessions. This increase was stronger in the group that had received the training for central anomalies first. For peripheral anomalies, there was no increase in false positive markings. The number of false positives was not addressed in Study 3, because we had no specific hypothesis regarding the occurrence of false positive markings. Taken together, the number of false positives regarding central anomalies was higher than for peripheral anomalies (Study 1) and increased due to the interventions (Studies 1 and 2).

To gain more insight into the underlying problems of OPT processing, explorative analyses on the proportion of different error types (H1c) were run for all three studies. Study 1 and Study 3 descriptively compared the frequency of bottom-up (detection errors) to top-down (recognition and decision) errors. The results revealed that more than 80% of the errors were top-down errors. In Study 3, the ratio of top-down to bottom-up errors did not seem to be influenced by the training. Similarly, the ratio of top-down to bottom-up errors was not affected by the training of Study 2. However, after the first training, students committed more top-down errors for central anomalies than before, or after the second training. A possible reason for this could be the different difficulties of the test sets to recognize the anomalies.

The following part reviews the results of the gaze behavior. The visual coverage was used as a measure in Studies 1 and 3. The gaze measures on anomalies (fixation time, number of fixations, and time to first fixation) were analyzed in all three studies.

The visual coverage (H2) was only investigated in Studies 1 and 3 because it was expected that only the training of search strategies would lead to higher visual coverage. In Study 1, the results showed a small effect of the intervention on visual coverage. The descriptive values of

Study 3 also indicated a small difference for visual coverage in favor of the intervention group. However, the results were not significant. Thus, the training methods did not have a major impact on visual coverage.

Contrary to expectations, none of the training increased the fixation time on anomalies (H3). In Study 1, the fixation time rather decreased in the intervention group and increased in the control group. A similar pattern was found in Study 3, where students in the intervention group also looked at anomalies for a shorter time than in the control group. The fixation time on anomalies was not affected by the training of Study 2. In general, students attended longer to peripheral than central anomalies in Study 1. Also in Study 2, different patterns for central and peripheral anomalies seemed to occur. The differentiation between peripheral and central anomalies was not made for Study 3.

Moreover, unexpectedly, no increases for the number of fixations on anomalies (H4) were found in the studies. Instead, in Study 1, students' number of fixations decreased in the intervention group and increased in the control group, especially for peripheral anomalies. Study 3 found the same effect in that there were fewer fixations on anomalies in the intervention group compared to control group. In Study 2, the pattern differed for peripheral and central anomalies with an increased number of fixations on peripheral anomalies after the second training. However, post-hoc comparisons did not find any effect on the number of fixations.

Finally, it was expected that the training would lead to shorter times to first fixation on anomalies (H5). In Study 1, the results showed the expected pattern that indicate a shift of attention towards peripheral anomalies and away from central anomalies. Thus, later fixations on central anomalies in the intervention group were found and earlier fixations in the control group. The pattern tended to be reversed for peripheral anomalies. The training of Study 2 did not influence the time to first fixation either for central or for peripheral anomalies. Contrary to expectations, in Study 3, students in the intervention group looked later at anomalies than students in the control group. However, the analysis in Study 3 did not differentiate between peripheral and central anomalies and thus the results are not directly comparable to the results of Study 1.

10.2 How to improve diagnostic performance in medical image processing?

The main goal of this dissertation is to develop and evaluate training methods that support dental students in their visual search and improve diagnostic performance. All studies were designed with a focus on improving the detection rate and thus on reducing false negative errors. Only the training evaluated in Study 2 achieved improvements in anomaly detection, whereas

the training from Study 1 or Study 3 showed no meaningful or no improvement. In a first step, I would like to discuss why only the training of Study 2 helped the students to detect anomalies. In a second step, the increase of false positive errors that occurred in Studies 1 and 2 is explored.

A first reason as to why only Study 2 supported anomaly detection may be the types of errors students make when interpreting OPTs. As investigated in Studies 1 and 3, the highest proportion of errors are top-down errors (recognition and decision errors) with more than 80 percent of all errors. These errors rely on top-down processing which means that they result from lacking knowledge about anatomical structures or pathological features (recognition errors) or because the observer decides against the relevance of recognized suspicious features (decision errors) (Brunyé et al., 2019). Thus, higher-level processes of working memory in interaction with information from long-term memory as described in Jarodzka, Boshuizen, et al. (2013) appear to play a crucial role in OPT processing. The training in Study 2 aimed to address exactly these top-down errors by providing information on visual features of anomalies and anatomical structures as well as by promoting the identification/categorization of anomalies with comparisons of radiographs. Hammer et al. (2008) claimed that instructional support is important for comparisons of different cases. The results of Study 2, which used highlighting of anatomical structures and pathological areas, support this assumption when considering that in former studies comparison without highlighting restricted the improvement of diagnostic performance to specific diseases (Kok et al., 2013) or were only efficient but not effective in detecting anomalies (Kok et al., 2015).

When starting with Study 1, a higher proportion of bottom-up (search) errors was expected considering the literature on other radiograph studies (Donovan & Litchfield, 2013; Kundel et al., 1978; Manning et al., 2004). Therefore, Study 1 focused on the reduction of bottom-up errors, which, in retrospect, account for only a small part of the errors typical for OPT processing in students. This might be the reason why the training of Study 1 only affected the detection of anomalies to a very small extent. In Study 3, the EMME training was expected to address both bottom-up and top-down errors. However, the training may not have been effective for other reasons such as a lack of necessary background knowledge or cognitive overload from the information-saturated EMME videos.

Second, the training in the three studies also differed in the degree to which they aimed at fostering either visual search strategies or object identification. While the training in Study 2 focused on object identification, especially Study 1 but also Study 3 evaluated search strategy training. Training object identification should be effective when searching for specific targets (Kramer et al., 2019). In Study 2, positive effects of such specific training for anomaly features

can be seen. Contrary, the training in Studies 1 and 3 took a holistic approach, targeting many types of anomalies with different visual characteristics (all anomalies in Study 3, all peripheral anomalies in Study 1). As mentioned by Kramer et al. (2019) object identification training aims at short-term effects, whereas search strategy training should provide long-term changes. As in Studies 1 and 3 only short-term effects were measured directly after the training in all studies, it is possible that potential effects of the training in Studies 1 and 3 would have only been seen in long-term measures. An indication that the two training methods could lead to long-term changes can be found in the gaze behavior. In Studies 1 and 3, the training led to similar changes in eye movements, showing that visual search was affected by the interventions, whereas in Study 2 there were almost no changes in gaze behavior. This change in Studies 1 and 3 could be a first step towards revealing later improvements in diagnostic performance. However, former research on gaze feedback, which was applied in Studies 1 and 3, also found beneficial short-term effects on diagnostic performance (Donovan et al., 2008; Gegenfurtner, Lehtinen, et al., 2017; Litchfield et al., 2010; Seppänen & Gegenfurtner, 2012). On the whole, the evidence is not yet clear and further studies are needed to examine the short- and long-term effects of search strategy training.

The distinction between object identification or search strategy training can be considered within the broader framework of problem-based learning concerning clinical reasoning. Here, researchers have disputed whether teaching knowledge or teaching the problem-solving process is more beneficial to medical students (Schmidt & Mamede, 2015). Training visual search strategies would belong to teaching problem-solving processes as search strategies cover the process of identifying relevant information in radiographs. In contrast, object identification training conveys background knowledge, for example, regarding the visual features of anomalies, the underlying mechanisms of diseases or anatomical structures of the radiograph. First evaluations indicate that teaching knowledge in clinical reasoning is more effective than teaching a problem-solving process (Monteiro et al., 2020; Schmidt & Mamede, 2015). Consistent with these results, the results of this dissertation found positive effects on anomaly detection only in Study 2, which corresponds to the teaching of knowledge, but not in Studies 1 and 3, which correspond mainly to the teaching of a problem-solving process. Thus, training aimed at fostering medical image processing should focus on teaching knowledge about what anomalies look like rather than just teaching viewing strategies (Kok & Jarodzka, 2017b).

The following section is devoted to the increase in false positive findings. Adequate diagnostic performance also includes observers identifying only the anomalies that are present and not detecting seemingly additional ones (false positive errors). False positives can still be

corrected retrospectively in OPT reporting, as the interpretation of the radiographs is only one part of the diagnostic process. For this reason, the primary goal of my work is not to reduce false positive errors, but to look at them exploratively to gain further insight into the visual processing skills dental students develop. The results from Studies 1 and 2 suggest that training increases the number of false positive errors, which seems to affect central anomalies in particular. In general, it is a common pattern that interventions lead to more false positive errors (Ganesan et al., 2018). Searching for anomalies in specific areas (similar to Study 1) or for specific anomalies (as in Study 2) has been shown to increase the number of anomalies also in other studies (Swensson et al., 1977, 1985). Possible explanations for these increases are that interventions interfere with the regular search behavior and the students might be encouraged by the intervention to find "as many anomalies as they can", which eventually leads to a higher possibility of also marking ambivalent areas. Besides, the artificial setting can lead to more false positive errors because students are not confronted with real patients and might not consider the consequences of false positive errors (e.g., subjecting the patient to additional, but unnecessary diagnostic procedures). Furthermore, the given search time, which some students experienced as overly long, especially in combination with lower confidence, could increase the chance to commit false positive errors (Ganesan et al., 2018). The fact that the number of false positive errors was higher for central than peripheral anomalies could be due to different prevalence (see above). Thus, greater familiarity with central anomalies could lead to an overestimation of students' ability to recognize central anomalies, which in turn leads to more errors (Berner & Graber, 2008). While in Study 1 both peripheral and central anomalies were affected, in Study 2 the number of false positives increased only for central anomalies. The reason for this pattern might be of technical nature. Study 2 addressed three peripheral and three central anomalies. Three types of peripheral anomalies cover a higher proportion of all possible anomalies in the periphery than three types of central anomalies with respect to all possible anomalies in the oral cavity. Students learned about only a small number of possible anomalies in the central area. Thus, students would later generally be more confident in dealing with the possible anomalies in the periphery than in the center and thus commit more false positive errors regarding central anomalies. One starting point to reduce false positives could be to make people more aware of the consequences of diagnostic errors.

To conclude, the training focusing on anomaly identification had a more positive effect on diagnostic performance than the training focusing on search strategies. Therefore, acquiring more background knowledge about diseases, characteristic features of anomalies, or anatomical structures is particularly relevant considering the high number of top-down errors. Thus, it

seems to be important to structure the university curricula of dental students in such a way that sufficient background knowledge is acquired before learning how to interpret radiographs. More training methods and evaluations are needed to further verify and specify the components of the training and also to prevent false positives.

10.3 Training effects regarding visual search

Across the three studies, it was assumed that the visual search for anomalies and the performance of detecting anomalies would be related to each other. Accordingly, the eye movements that reflect the visual search should also change as a result of the training in all three studies that aimed to improve the interpretation of medical images and the visual search. It is important to note here that Study 2, with object identification training, should also improve visual search. If an intervention were to train pathological background knowledge or basic science, it would not be reasonable to assume large changes in visual search, since only higher order cognitive processes are addressed (Jarodzka, Boshuizen, et al., 2013). However, the training in Study 2 focused on teaching the visual appearance of anomalies' characteristic visual features, in which knowledge-driven object identification is related to general perceptual processes of visual search.

In Studies 1 and 3, which directly addressed the visual search strategy, there should have been an increase in the visual coverage rate of OPTs in dental students. However, a small increase was only found in Study 1, but not in Study 3. One explanation could be that in Study 1 the goal was to achieve full coverage of OPTs with the individualized gaze feedback, whereas in Study 3 full coverage was not explicitly emphasized in the EMME videos. Although the expert models in the EMME videos provided didactical verbal explanations simultaneously and used a didactic approach to visual search that emphasized attending to all areas of an OPT, It could be that students did not perceive the EMME videos to focus on full coverage, so students did not apply full coverage to their own visual search behavior.

The training in all three studies focused on anomaly detection. Thus, it is of particular interest to see how gaze behavior changed regarding the anomalies. The training should have led to a more intense visual processing of addressed anomalies reflected in longer and more fixations derived from the first study of the overall project (Richter et al., 2020). However, the opposite pattern was found in Study 1 and 3. Here the training led to shorter and fewer fixations on anomalies indicating a more efficient visual search (as the same level of accuracy was achieved with less processing). So far, little is known on the development of visual search within students because most previous research compared experts with novices rather than

applying a more-fined approach that differentiates within intermediate levels of performance (Gegenfurtner et al., 2011). Experts seem to make less fixations in general on medical images (van der Gijp et al., 2017) and also fewer fixations on anomalies for OPTs (Grünheid et al., 2013; Turgeon & Lam, 2016). Thus, the fewer fixations found in Studies 1 and 3 may also reflect that the training enabled students to take their next step in developing their medical image processing. Additionally, the fewer and shorter fixations on anomalies may suggest that students applied more expanded visual processing. Thus, the visual search training would lead the students to not only focus on the anomalies but also on other areas. Thus, this gaze behavior would indicate a global rather than a focal search when interpreted against the backdrop of the global-focal search model (cf. van der Gijp et al., 2017). Perhaps the training improved the global processing aspect, which is typically not that well established in students, who normally use a search-to-find method (Kundel et al., 2007; Nodine & Mello-Thoms, 2000). This would also indicate that students were developing their image processing to a higher level. However, necessary background knowledge might still be lacking to really benefit from this development of visual search in terms of diagnostic performance (Kok et al., 2012).

For the time to first fixate an anomaly, Studies 1 and 3 showed different results. In line with the expectations, students in the intervention group of Study 1 looked later at central anomalies and tended to look sooner at peripheral anomalies after the intervention. Contrary to expectations, students in the intervention group of Study 3, however, looked later at the anomalies in general. A reason for these differences between Studies 1 and 3 might be that Study 3 did not differentiate the location of anomalies any further. When considering the location of anomalies in the analysis[3] the same pattern as in Study 1 results with later fixations on central anomalies and a tendency for sooner fixations on peripheral anomalies. Thus, it may be that the training shifted the focus of attention away from the oral cavity to the periphery with low-prevalence anomalies that were less familiar than central anomalies. The time to first fixation is typically seen as an indicator of global processing, as derived from studies on mammography processing (Kundel et al., 2007). Thus, early attention to anomalies is

[3]This analysis was in addition to the analyses in the paper of Study 3.

The time to first fixation on anomalies for peripheral and central anomalies was analyzed with the same linear mixed models as in Study 1 and 2. The results showed a significant interaction between time, condition, and location (χ^2 (1) = 9.59, p = .002), meaning that the EMME training influenced the time to first fixation differently for central and peripheral anomalies. Post-hoc pairwise comparisons showed that in the intervention group the time to first fixation on central anomalies increased from the pre- (M = 27339 *ms*, SD = 22667 *ms*) to the post-test (M = 35464 *ms*, SD = 23972 *ms*), Estimate = 8404.00, z = 3.46, p = .006. For peripheral anomalies the descriptive value of pre- (M = 23497 *ms* , SD = 20246 *ms*) and post-test (M = 22372 *ms* , SD = 22239 *ms*) indicate a small decrease. However, the post-hoc comparisons were not significant, Estimate = -1844.8, z = -.45, p = 1.00. In the control group, the time to first fixation did not change from pre-to post-test, all p > .05.

interpreted as an expert's profound ability to catch an anomaly at first glimpse. However, even though time to first fixation has also revealed expert-novice differences in OPT processing (Turgeon & Lam, 2016), there is a concern that this measure may be less informative for OPTs. OTPs, due to the large number of anomalies contained therein, require hybrid search (Wolfe, 2012; Wolfe et al., 2016). Thus, students might see some anomalies right at the beginning but detect others later. The resulting mean would indicate an averaged time to first fixation, which is not very informative and occludes differences in visual search. Importantly, the OPTs used in the present study contained more anomalies than those used by Turgeon and Lam (2016), where the OPTs contained one to four anomalies only. This can explain why Turgeon and Lam (2016) found meaningful effects corresponding to results from radiography studies in other domains despite using OPTs.

The results of Study 2 indicate that the training with radiograph comparison did not affect the students' gaze behavior regarding anomalies. These results are unexpected because the training in this study led to an improvement in the detection of anomalies and it is precisely then that changes in eye movements should become apparent. A possible reason for this might be the type of training, which aimed at anomaly identification and reduction of top-down errors but did not teach viewing strategies as was the case in the training in Studies 1 and 3.Thus, students might improve their cognitive processes such as being able to retrieve relevant knowledge information for anomaly identification. Such effects at a cognitive level would likely be reflected in other eye movements such as pupil dilation, which is a measure for cognitive load and which has been shown to be informative in medical image processing (Brunyé et al., 2016; Castner et al., 2020). Further studies are needed to investigate these assumptions.

To summarize, training affected gaze behavior during visual search when it addressed visual search strategies. Gaze behavior appeared to change toward more efficient and global visual search with fewer and shorter fixations. When the training focused on anomaly identification instead, no changes in gaze behavior were observed. On the whole, gaze behavior contributed relatively little to explaining findings regarding OPT interpretation in contrast to what has been found in other medical image processing studies (e.g., Jaarsma et al., 2014; Kundel et al., 2007; Manning et al., 2006).

10.4 Strengths and limitations

This dissertation has strengths and limitations regarding the studies that were conducted. The strengths of the present work relate to the use of a use-inspired basic research approach, its methodology and analyses.

First, this dissertation implemented a use-inspired basic research program, which is based on theoretical assumptions and empirical evidence and provides applications for real university teaching. Realistic materials in form of OPTs were used, which, depending on their occurrence in the field, have different properties. Also, the developed training methods were evaluated with a specialized population, namely, dental students from different semesters who would also be recipients of such training in university teaching. Thus, the evaluation of the developed training methods has a high ecological validity and could be applied to university teaching without transformation. Of course, this is only useful if the training method is also effective which is the case for the training of Study 2. In particular, this training could be easily implemented into curricula as it does not require special technical equipment. Furthermore, this dissertation integrated the dental and radiographic expertise of two domain experts (maxillofacial radiologist and a prosthodontist) into psychological research on medical image processing. This cooperation between psychology and the medical domain made it possible to conduct this use-inspired basic research that can be applied in medical education.

Second, in this dissertation different training methods to identify effective methods for OPT interpretation were evaluated. The training methods themselves provide methodological variety with individualized or instructional methods combined with compare-and-contrast tasks or modeling examples. To the best of my knowledge, this research is one of the first that developed and evaluated an individualized training intervention in this field (Kok et al., 2017). Additionally, in Studies 1 and 3 innovative methods with eye movements as supportive tool were applied. Besides, eye movements were used to measure changes in students' visual search to gain insight into their perceptual and cognitive processes. Especially when bringing eye movements and diagnostic performance together, detailed insight into the diagnostic errors was examined, which is very important to develop applied training methods. Thus, this dissertation combined different methodologies such as eye tracking and diagnostic performance which is important to gain knowledge about visual image processing (Jarodzka & Boshuizen, 2017).

Third, state-of-the-art statistical analyses to evaluate the training methods were used. The complex designs in Studies 1 and 2 with different measurement times and groups of addressed anomalies as well as the individual characteristics of the OPTs with different degrees of difficulty to recognize anomalies required complex statistical evaluation methods such as

(generalized) linear mixed models. On the whole, this work is characterized by the fact that it conducts application-based research that is transferable to practice, brings together a variety of methods and evaluates them with precise and complex statistical procedures.

Nevertheless, especially the applied nature of this research also yields limitations concerning the lack of experimental control due to the use of naturally occurring stimuli, the procedure and the special target population.

First, the occurrence of anomalies, which are differentially difficult to detect, and the combination of anomalies within one OPT is not controllable when using naturally occurring materials. Thus, it is not possible to completely counterbalance, for example, the number of anomalies for different test sets or the level of difficulty of the test sets. This fact may have limited the validity of the study results when using different OPTs for the test sets at different times of measurement as was the case in Studies 2 and 3. Furthermore, there may be dependencies between anomalies or different types of anomalies. When a student detected, for example, an apical radiolucency it might be reasonable that s/he searched for further anomalies of apical radiolucency and thereby missed other anomalies adjacent to the first apical radiolucency. Also, it could be possible that finding an apical radiolucency leads to more findings of insufficient root fillings, which can cause apical radiolucency. If students have already developed such causal concepts that drive their visual search, it would be difficult and not necessarily desirable for such brief training interventions to interfere with them. Such dependencies could be especially important when searching for multiple anomalies, which is the case for OPTs, because exhaustion of working memory capacity could stop the search before all anomalies are detected (Brunyé et al., 2019). So far, to my knowledge, there are no studies that have investigated such dependencies between anomalies in OPTs.

Second, the students saw the OPTs twice in a search and a marking phase. Students' gaze behavior was only analyzed during the search phase, in which the students only had to search for anomalies. In the marking phase, students only had to mark the anomalies, but it cannot completely rule out that some students might have continued their search for anomalies. Therefore, the results regarding eye movements might be biased and their value for basic science might be limited. However, the intention of this design was to separate these two phases to avoid possible interference effects due to the action of marking the anomaly with the mouse, which would also have affected the eye movements during visual search (e.g., looking at the mouse, following the marking tool with the eyes). Especially when considering that the OPTs show many anomalies, the action of marking would disturb the visual search processes. Thus, integrating both phases into one was not a viable alternative.

Third, this work is based on a special population, namely dental students. This brings along two limitations: On the one hand, it is not clear if the results are also generalizable for dentists with higher image processing skills. It might also be interesting to provide training for experienced dentists because they still need further training to guarantee high performance (Ericsson, 2004). On the other hand, a special population makes it difficult to conduct studies, as only a small number of subjects are available. Therefore, it was not possible to examine all components of the training separately to determine the extent to which they were effective or whether they were effective only when the components were combined, such as verbal explanations, highlighting, and comparisons in Study 2 or the three different EMME videos and the didactical verbal explanations of Study 3. However, from a basic science perspective, the studies provide a very interesting starting point for further research to evaluate the single components and their effects.

10.5 Implications and further directions

This section describes the theoretical and practical implications of this work and states what future research ideas can be derived from it.

The exploratory analyses in all three studies provide information on the distribution of detection, recognition and decision errors. Detection errors appear to play only a minor role in the interpretation of OPTs with less than 20%. This proportion is thus lower on average than for chest radiographs (Donovan & Litchfield, 2013; Kundel et al., 1978; Manning et al., 2004). This observation supports the assumption that the processing of medical images is task-related and domain-specific (Gegenfurtner et al., 2011; van der Gijp et al., 2017). As the threshold to differentiate between recognition and decision error seems to depend on different characteristics of radiographs (Brunyé et al., 2019) and no studies have investigated these error types in OPTs, this dissertation does not differentiate between recognition and decision errors in the analysis. For more basic research, it would be interesting to investigate how to distinguish between recognition and decision errors in OPTs, for example, based on pupil dilation (cf. Brunyé et al., 2019). Such differentiation could give further insight into the problems when interpreting OPTs and thus be helpful in developing further training methods.

Furthermore, the question arises to what extent the global focal models are applicable to OPT interpretation. Originally, the global focal models were derived from mammography and chest radiographs with mostly single anomalies (Kundel et al., 2007; Nodine & Mello-Thoms, 2000). In OPTs, an observer needs to perform a hybrid search for many anomalies and it is not clear if the hybrid search underlies a global and focal search (Wolfe, 2012). This work can

provide tentative insight into this question. As the training in Studies 1 and 2 addressed anomalies located either in the center or the periphery of the OPT in different ways, it could be deduced that the visual search also differs for these areas. Thus, there appears to be no overall mechanism of visual search that covers all areas of the OPT at once as would be the case in a global search. Besides, it is questionable if a first global impression is helpful when multiple anomalies are present. Can the first global impression capture multiple suspicious regions or does the global impression end when finding one suspicious region that is further inspected in a focal search? In Study 2, which found supportive effects for the detection of anomalies, a change of visual search would be expected. A change towards a global processing should then be reflected in shorter time to first fixations. However, no effects on the time to first fixate an anomaly could be found. Therefore, it does not appear that the global focal model can be applied to hybrid search. The results for the other eye tracking parameters, most of which showed no changes in Study 2, also indicate that hybrid search is different from the search for individual anomalies, such as in mammography.

Still, further research with experts, who contrary to novices use such global impression, is needed to verify these conclusions and answer these questions of visual search in OPTs. Former research already indicates that global search is not only related to the first fixation but potentially includes peripheral vision, which helps to process multiple areas in parallel (Litchfield & Donovan, 2016; Sheridan & Reingold, 2017). Thus, processing of OPTs, which contain many anomalies, would be an interesting research object to further verify this observation.

From a practical point of view, the results of this work represent another step towards the development of effective training methods. The evaluation of the training focusing on full coverage and reduction of detection errors in Study 1 verified the results of previous studies, which showed for chest radiographs that full coverage or systematic viewing did not improve diagnostic performance (Kok et al., 2016; van Geel et al., 2017). Thus, only teaching systematic viewing strategies and full coverage search is not beneficial for medical students in terms of obtaining better diagnostic performance (Waite et al., 2019). Instead, providing knowledge about what anomalies look like as the training in Study 2 did with multiple methods seems to be a more promising approach (Kok & Jarodzka, 2017b). As this training method with comparison of radiographs, highlighting of anatomical structures and pathological areas, and descriptions about the visual appearance of anomalies does not need elaborate technical equipment, it would be easy to apply these methods in university teaching. Further research could evaluate whether the positive effects of this training are also generalizable to other

medical images, for instance, chest radiographs. The practical relevance of EMME for the interpretation of OPTs could not be supported in Study 3. Thus, research which further investigates EMME and considers, for example, potential cognitive overload due to the EMME videos is needed before putting this method into practice. Furthermore, it seems to be very important to investigate the nature of the errors and the cognitive problems before developing and applying training. Only when addressing the underlying problems of medical image processing could training be effective and used in practice.

In general, Study 2, which focused on the identification of objects, was more beneficial than Studies 1 and 3, which used viewing strategy training. Kramer et al. (2019) assume that object identification training shows short-term effects whereas viewing strategy training shows long-term effects. Since Studies 1 and 3, which used search strategy training, only measured short-term effects, this could be the reason why there were no positive effects on performance. From this perspective, it would be very interesting to study the long-term effects of these training methods in future studies. The change of eye movements that could be observed in Studies 1 and 3 might be a first indicator that students cognition changed which could potentially lead to long-term changes in diagnostic performance.

Based on the results of this dissertation, it is questionable whether the radiology course for learning visual search by interpreting dental images is well prescribed in the university curriculum. Currently, the course is held in the 6th semester and mainly addresses the process of visual search with search strategies, massed practice with 100 radiographs and students receive technical information on imaging in this course (Richter et al., 2020). Before the 6th semester, students take the preliminary medical examination, which covers the basics of medicine, but they only learn little dental content. Considering that knowledge of dental anomalies, dental anatomical structures, and dental pathologies is important in identifying anomalies in dental radiographs, the radiology course should be taken rather later in the curriculum when sufficient knowledge is already available. This view is supported by the prerequisites for medical image processing of van der Gijp et al. (2014). The authors also stated that knowledge of anatomy, pathology, radiological image techniques is important for all the three cognitive processes – Perception, Synthesis and Analysis - involved in medical image interpretation.

Furthermore, the results of this work have great potential for practical application. They could be used to design, for example, an online learning environment for medical image interpretation. Comparing radiographs with highlights could be a component that might be

augmented by individual feedback on learner's performance and biomedical knowledge about anomalies (cf. Baghdady et al., 2009; Ericsson, 2015).

10.6 Conclusion

At a global level, this dissertation confirms that psychology can substantially contribute to the medical field of clinical reasoning. Complex visual processing of medical images that required high-level performance were studied with psychological methods. Psychological research contributed to design medical training and the evaluation of this training showed that knowledge plays a crucial role when interpreting medical images.

At a more detailed level, the purpose of this dissertation was to develop and evaluate training interventions to assist dental students in achieving high skills in medical image processing. In doing so, the three training methods focused on reducing the error-prone medical image processing process related to visual search and image interpretation that occurs when dental students evaluate panoramic radiographs. The training included a wide range of different intervention methods such as individualized gaze feedback in a compare-and-contrast task, comparisons of radiographs with different highlighting of relevant areas, and eye movement modeling examples. The results showed that only the method of comparing radiographs with different highlighting of relevant areas supported the diagnostic performance, whereas EMME and the individualized gaze feedback did not show positive effects. Further studies are needed to investigate if these training methods show potentially long-term effects. Nevertheless, the two later training methods which focused on viewing strategies changed the eye movements of students to a more expanded visual search. In addition, this work showed that dental students' processing is affected less by detection errors than it is by recognition and decision errors. This could be the reason why only the training intervention that directly tried to reduce recognition and decision errors by providing missing background knowledge about the visual characteristics of anomalies was effective. For university teaching it is important to provide the basic requirements and teach anatomical and pathological knowledge and specifically the visual occurrence of anomalies rather than teaching systematic visual search strategies (Kok & Jarodzka, 2017b; van der Gijp et al., 2014). This work substantiates the importance of developing training methods according to the needs of learners in order to improve the diagnostic decisions made by physicians, thus improving the well-being of patients and, in severe cases, saving their lives.

11. References

Al-Moteri, M. O., Symmons, M., Plummer, V., & Cooper, S. (2017). Eye tracking to investigate cue processing in medical decision-making: A scoping review. *Computers in Human Behavior*, *66*, 52–66. https://doi.org/10.1016/j.chb.2016.09.022

AMBOSS GmbH. (2019). *Silhuettenphänomen auf einer Röntgen-Thorax-Aufnahme*. https://www.amboss.com/de/wissen/Befundung_eines_Röntgen-Thorax

Baddeley, A. (1992). Working memory. *Science*, *255*(5044), 556–559. https://doi.org/10.1126/science.1736359

Baghdady, M., Carnahan, H., Lam, E. W. N., & Woods, N. N. (2013). Integration of basic sciences and clinical sciences in oral radiology education for dental students. *Journal of Dental Education*, *77*(6), 757–763. https://doi.org/10.1002/j.0022-0337.2013.77.6.tb05527.x

Baghdady, M., Carnahan, H., Lam, E. W. N., & Woods, N. N. (2014). Test-enhanced learning and its effect on comprehension and diagnostic accuracy. *Medical Education*, *48*(2), 181–188. https://doi.org/10.1111/medu.12302

Baghdady, M., Pharoah, M. J., Regehr, G., Lam, E. W. N., & Woods, N. N. (2009). The role of basic sciences in diagnostic oral radiology. *Journal of Dental Education*, *73*(10), 1187–1193. https://doi.org/10.1002/j.0022-0337.2009.73.10.tb04810.x

Bahaziq, A., Jadu, F. M., Jan, A. M., Baghdady, M., & Feteih, R. M. (2019). A comparative study of the examination pattern of panoramic radiographs using eye-tracking software. *Journal of Contemporary Dental Practice*, *20*(12), 1436–1441. https://doi.org/10.5005/jp-journals-10024-2700

Bandura, A. (1971). *Social learning theory*. General Learning Press.

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Berner, E. S., & Graber, M. L. (2008). Overconfidence as a cause of diagnostic error in medicine. *American Journal of Medicine*, *121*(5A), 2–23. https://doi.org/10.1016/j.amjmed.2008.01.001

Boshuizen, H. P. A., Gruber, H., & Strasser, J. (2020). Knowledge restructuring through case processing: The key to generalise expertise development theory across domains? *Educational Research Review*, *29*, Article 100310. https://doi.org/10.1016/j.edurev.2020.100310

Boshuizen, H. P. A., & Schmidt, H. G. (1992). On the role of biomedical knowledge in clinical reasoning by experts, intermediates and novices. *Cognitive Science*, *16*(2), 153–184. https://doi.org/10.1016/0364-0213(92)90022-M

Boshuizen, H. P. A., & Schmidt, H. G. (2008). The development of clinical reasoning expertise. In J. Higgs, M. A. Jones, S. Loftus, & N. Christensen (Eds.), *Clinical Reasoning in the Health Professions* (3rd ed., pp. 113–121). Elsevier Health Science.

Brunyé, T. T., Drew, T., Weaver, D. L., & Elmore, J. G. (2019). A review of eye tracking for understanding and improving diagnostic interpretation. *Cognitive Research: Principles and Implications*, *4*, Article 7. https://doi.org/10.1186/s41235-019-0159-2

Brunyé, T. T., Eddy, M. D., Mercan, E., Allison, K. H., Weaver, D. L., & Elmore, J. G. (2016). Pupil diameter changes reflect difficulty and diagnostic accuracy during medical image interpretation. *BMC Medical Informatics and Decision Making*, *16*(1), 1–8. https://doi.org/10.1186/s12911-016-0322-3

Bundesamt für Strahlenschutz. (2016). *Röntgendiagnostik: Häufigkeit und Strahlenexposition*. http://www.bfs.de/DE/themen/ion/anwendung-medizin/diagnostik/roentgen/haeufigkeit-exposition.html

Carmody, D. P., Kundel, H. L., & Toto, L. C. (1984). Comparison scans while reading chest images: taught, but not practiced. *Investigative Radiology*, *19*, 462–466. https://doi.org/10.1097/00004424-198409000-00023

Castner, N., Appel, T., Eder, T., Richter, J., Scheiter, K., Keutel, C., Hüttig, F., Duchowski, A., & Kasneci, E. (2020). Pupil diameter differentiates expertise in dental radiography visual search. *PLoS ONE*, *15*(5), Article e0223941. https://doi.org/10.1371/journal.pone.0223941

Castner, N., Kasneci, E., Kübler, T., Scheiter, K., Richter, J., Eder, T., Hüttig, F., & Keutel, C. (2018). Scanpath comparison in medical image reading skills of dental students. *Eye Tracking Research and Applications Symposium (ETRA)*. https://doi.org/10.1145/3204493.3204550

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates.

Collins, A., & Kapur, M. (2014). Cognitive apprenticeship. In R. Sawyer (Ed.), *The Cambridge Handbook of the Learning Sciences* (pp. 109–127). Cambridge University Press. https://doi.org/10.1080/14739879.2015.1101851

Constantine, S., Roach, D., Liberali, S., Kiermeier, A., Sarkar, P., Jannes, J., Sambrook, P., Anderson, P., & Beltrame, J. (2018). Carotid artery calcification on Orthopantomograms (CACO Study) – is it indicative of carotid stenosis? *Australian Dental Journal*, *64*, 4–10. https://doi.org/10.1111/adj.12651

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

Donovan, T., & Litchfield, D. (2013). Looking for cancer: Expertise related differences in searching and decision making. *Applied Cognitive Psychology*, *27*(1), 43–49. https://doi.org/10.1002/acp.2869

Donovan, T., Manning, D. J., & Crawford, T. (2008). Performance changes in lung nodule detection following perceptual feedback of eye movements. *Proceedings of SPIE, Medical Imaging 2008: Image Perception, Observer Performance, and Technology Assessment*, *691703*. https://doi.org/10.1117/12.768503

Drew, T., & Williams, L. H. (2017). Simple eye-movement feedback during visual search is not helpful. *Cognitive Research: Principles and Implications*, *2*(1), 44. https://doi.org/10.1186/s41235-017-0082-3

Eder, T. F., Richter, J., Scheiter, K., Huettig, F., & Keutel, C. (2021). Comparing radiographs with signaling improves anomaly detection of dental students: An eye-tracking study. *Applied Cognitive Psychology*, 1–15. https://doi.org/10.1002/acp.3819

Eder, T. F., Richter, J., Scheiter, K., Keutel, C., Castner, N., Kasneci, E., & Huettig, F. (2021). How to support dental students in reading radiographs : effects of a gaze - based compare - and - contrast intervention. *Advances in Health Sciences Education*, *26*(1), 159–181. https://doi.org/10.1007/s10459-020-09975-w

Emhardt, S. N., Kok, E. M., Jarodzka, H., Brand-Gruwel, S., Drumm, C., & van Gog, T. (2020). How experts adapt their gaze behavior when modeling a task to novices. *Cognitive Science*, *44*(9), Article e12893. https://doi.org/10.1111/cogs.12893

Ericsson, K. A. (2004). Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. *Academic Medicine*, *79*(10), 70–81. https://doi.org/10.1097/00001888-200410001-00022

Ericsson, K. A. (2015). Acquisition and maintenance of medical expertise: A perspective from the expert-performance approach with deliberate practice. *Academic Medicine*, *90*(11), 1471–1486. https://doi.org/10.1097/ACM.0000000000000939

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, *100*(3), 363–406. https://doi.org/10.1037/0033-295X.100.3.363

Fawver, B., Thomas, J. L., Drew, T., Mills, M. K., Auffermann, W. F., Lohse, K. R., & Williams, A. M. (2020). Seeing isn't necessarily believing: Misleading contextual information influences perceptual-cognitive bias in radiologists. *Journal of Experimental Psychology: Applied*. https://doi.org/10.1037/xap0000274

Friedlander, A. H., & Freymiller, E. G. (2003). Detection of radiation-accelerated atherosclerosis of the carotid artery by panoramic radiography. A new opportunity for dentists. *The Journal of the American Dental Association*, *134*(10), 1361–1365. https://doi.org/10.14219/jada.archive.2003.0052

Friedlander, A. H., Garrett, N. R., Chin, E. E., & Baker, J. D. (2005). Ultrasonographic confirmation of carotid artery atheromas diagnosed via panoramic radiography. *The Journal of the American Dental Association*, *136*(5), 633–635. https://doi.org/10.14219/jada.archive.2005.0235

Friedlander, A. H., Garrett, N. R., & Norman, D. C. (2002). The prevalence of calcified carotid artery atheromas on the panoramic radiographs of patients with type 2 diabetes mellitus. *The Journal of the American Dental Association*, *133*(11), 1516–1523. https://doi.org/10.14219/jada.archive.2002.0083

Ganesan, A., Alakhras, M., Brennan, P. C., & Mello-Thoms, C. (2018). A review of factors influencing radiologists' visual search behaviour. *Journal of Medical Imaging and Radiation Oncology*, *62*, 747–757. https://doi.org/10.1111/1754-9485.12798

Garland, L. H. (1949). On the scientific evaluation of diagnostic procedures. *Radiology*, *52*(3), 309–328. https://doi.org/10.1148/52.3.309

Geel, K. van, Kok, E. M., Aldekhayel, A. D., Robben, S. G. F., & van Merriënboer, J. J. G. (2018). Chest X-ray evaluation training: impact of normal and abnormal image ratio and instructional sequence. *Medical Education*, *53*(2), 153–164. https://doi.org/10.1111/medu.13756

Gegenfurtner, A., Kok, E., van Geel, K., de Bruin, A., Jarodzka, H., Szulewski, A., & van Merriënboer, J. J. G. (2017). The challenges of studying visual expertise in medical image diagnosis. *Medical Education*, *51*(1), 97–104. https://doi.org/10.1111/medu.13205

Gegenfurtner, A., Lehtinen, E., Jarodzka, H., & Säljö, R. (2017). Effects of eye movement modeling examples on adaptive expertise in medical image diagnosis. *Computers and Education*, *113*, 212–225. https://doi.org/10.1016/j.compedu.2017.06.001

Gegenfurtner, A., Lehtinen, E., & Säljö, R. (2011). Expertise differences in the comprehension of visualizations: A meta-analysis of eye-tracking research in professional domains. *Educational Psychology Review*, *23*, 523–552. https://doi.org/10.1007/s10648-011-9174-7

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology*, *106*(3), 608–626. https://doi.org/10.1037/a0034716

Grant, E. R., & Spivey, M. J. (2003). Eye movements and problem solving: Guiding attention guides thought. *Psychological Science*, *14*(5), 462–466. https://doi.org/10.1111/1467-9280.02454

Green, D. M., & Swets, J. A. (1988). *Signal detection theory and psychophysics*. Peninsula Publishing.

Grunewald, M., Heckemann, R., Gebhard, H., Lell, M., & Bautz, W. (2003). COMPARE Radiology: Creating an interactive web-based training program for radiology with multimedia authoring software. *Academic Radiology*, *10*(5), 543–553. https://doi.org/10.1016/S1076-6332(03)80065-X

Grünheid, T., Hollevoet, D. A., Miller, J. R., & Larson, B. E. (2013). Visual scan behavior of new and experienced clinicians assessing panoramic radiographs. *Journal of the World Federation of Orthodontists*, *2*(1), 3–7. https://doi.org/10.1016/j.ejwf.2012.12.002

Gruppen, L. D. (2017). Clinical reasoning: Defining it, teaching it, assessing it, studying it. *Western Journal of Emergency Medicine*, *18*(1), 4–7. https://doi.org/10.5811/westjem.2016.11.33191

Hammer, R., Bar-Hillel, A., Hertz, T., Weinshall, D., & Hochstein, S. (2008). Comparison processes in category learning: From theory to behavior. *Brain Research*, *1225*, 102–118. https://doi.org/10.1016/j.brainres.2008.04.079

Hammer, R., Diesendruck, G., Weinshall, D., & Hochstein, S. (2009). The development of category learning strategies: What makes the difference? *Cognition*, *112*(1), 105–119. https://doi.org/10.1016/j.cognition.2009.03.012

Hermanson, B. P., Burgdorf, G. C., Hatton, J. F., Speegle, D. M., & Woodmansey, K. F. (2018). Visual Fixation and Scan Patterns of Dentists Viewing Dental Periapical Radiographs: An Eye Tracking Pilot Study. *Journal of Endodontics*, *44*(5), 722–727. https://doi.org/10.1016/j.joen.2017.12.021

Hillard, A., Myles-Worsley, M., Johnston, W., & Baxter, B. (1985). The development of radiologic schemata through training and experience. *Investigative Radiology*, *4*, 422–425. https://doi.org/10.1097/00004424-198507000-00017

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 University Press.

Isaacs, E. A., & Clark, H. H. (1987). References in conversation between experts and novices. *Journal of Experimental Psychology: General*, *116*(1), 26–37. https://doi.org/10.1037/0096-3445.116.1.26

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

Jaarsma, T., Jarodzka, H., Nap, M., van Merriënboer, J. J. G., & Boshuizen, H. P. A. (2015). Expertise in clinical pathology: combining the visual and cognitive perspective. *Advances in Health Sciences Education*, *20*(4), 1089–1106. https://doi.org/10.1007/s10459-015-9589-x

Jarodzka, H., Balslev, T., Holmqvist, K., Nyström, M., Scheiter, K., Gerjets, P., & Eika, B. (2012). Conveying clinical reasoning based on visual observation via eye-movement modelling examples. *Instructional Science*, *40*(5), 813–827. https://doi.org/10.1007/s11251-012-9218-5

Jarodzka, H., & Boshuizen, H. P. A. (2017). Unboxing the black box of visual expertise in medicine. *Frontline Learning Research*, *5*(3), 167–183. https://doi.org/10.14786/flr.v5i3.322

Jarodzka, H., Boshuizen, H. P. A., & Kirschner, P. A. (2013). Cognitive skills in medicine. In P. Lanzer (Ed.), *Catheter-based cardiovascular interventions: a knowledge-based approach* (pp. 69–86). Springer Science & Business Media. https://doi.org/10.1007/978-3-642-27676-7

Jarodzka, H., van Gog, T., Dorr, M., Scheiter, K., & Gerjets, P. (2013). Learning to see: Guiding students' attention via a model's eye movements fosters learning. *Learning and Instruction*, *25*, 62–70. https://doi.org/10.1016/j.learninstruc.2012.11.004

Jensen, G., Resnik, L., & Haddad, A. (2008). Expertise and clinical reasoning. In J. Higgs, M. Jones, S. Loftus, & N. Christensen (Eds.), *Clinical Reasoning in the Health Professions* (3rd ed., pp. 123–136). Elsevier Health Science.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, *87*(4), 329–354. https://doi.org/10.1037/0033-295X.87.4.329

Kok, E. M. (2019). Eye tracking: the silver bullet of competency assessment in medical image interpretation? *Perspectives on Medical Education, 8*, 63-64. https://doi.org/10.1007/s40037-019-0506-5

Kok, E. M., de Bruin, A. B. H., Leppink, J., van Merriënboer, J. J. G., & Robben, S. G. F. (2015). Case comparisons: An efficient way of learning radiology. *Academic Radiology*, *22*(10), 1226–1235. https://doi.org/10.1016/j.acra.2015.04.012

Kok, E. M., de Bruin, A. B. H., Robben, S. G. F., & van Merriënboer, J. J. G. (2012). Looking in the same manner but seeing it differently: Bottom-up and expertise effects in radiology. *Applied Cognitive Psychology*, *26*(6), 854–862. https://doi.org/10.1002/acp.2886

Kok, E. M., de Bruin, A. B. H., Robben, S. G. F., & van Merriënboer, J. J. G. (2013). Learning radiological appearances of diseases: Does comparison help? *Learning and Instruction*, *23*, 90–97. https://doi.org/10.1016/j.learninstruc.2012.07.004

Kok, E. M., & Jarodzka, H. (2017a). Before your very eyes: The value and limitations of eye tracking in medical education. *Medical Education*, *51*(1), 114–122. https://doi.org/10.1111/medu.13066

Kok, E. M., & Jarodzka, H. (2017b). Beyond your very eyes: eye movements are necessary, not sufficient. *Medical Education*, *51*(11), 1190. https://doi.org/10.1111/medu.13384

Kok, E. M., Jarodzka, H., de Bruin, A. B. H., BinAmir, H. A. N., Robben, S. G. F., & van Merriënboer, J. J. G. (2016). Systematic viewing in radiology: seeing more, missing less? *Advances in Health Sciences Education*, *21*, 189–205. https://doi.org/10.1007/s10459-015-9624-y

Kok, E. M., van Geel, K., van Merriënboer, J. J. G., & Robben, S. G. F. (2017). What we do and do not know about teaching medical image interpretation. *Frontiers in Psychology*, *8*, Article 309. https://doi.org/10.3389/fpsyg.2017.00309

Kourdioukova, E. V., Valcke, M., Derese, A., & Verstraete, K. L. (2011). Analysis of radiology education in undergraduate medical doctors training in Europe. *European Journal of Radiology*, *78*(3), 309–318. https://doi.org/10.1016/j.ejrad.2010.08.026

Kramer, M. R., Porfido, C. L., & Mitroff, S. R. (2019). Evaluation of strategies to train visual search performance in professional populations. *Current Opinion in Psychology*, *29*, 113–118. https://doi.org/10.1016/j.copsyc.2019.01.001

Krebs, M. C., Schüler, A., & Scheiter, K. (2019). Just follow my eyes: The influence of model-observer similarity on eye movement modeling examples. *Learning and Instruction*, *61*, 126–137. https://doi.org/10.1016/j.learninstruc.2018.10.005

Kübler, T. C., Sippel, K., Fuhl, W., Schievelbein, G., Aufreiter, J., Rosenberg, R., Rosenstiel, W., & Kasneci, E. (2015). Analysis of eye movements with Eyetrace. *International Conference on Biomedical Engineering Systems and Technologies*, 458–471.

Kundel, H. L., & La Follette, P. S. (1972). Visual search patterns and experience with radiological images. *Diagnostic Radiology*, *103*(3), 523–528. https://doi.org/10.1148/103.3.523

Kundel, H. L., Nodine, C. F., & Carmody, D. (1978). Visual scanning, pattern recognition and decision-making in pulmonary nudle detection. *Investigative Radiology*, *13*(3), 175–181. https://doi.org/10.1097/00004424-197805000-00001

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

Kundel, H. L., Nodine, C. F., & Krupinski, E. A. (1990). Computer-displayed eye position as a visual aid to pulmonary nodule interpretation. *Investigative Radiology*, *25*(8), 890–896. https://doi.org/10.1097/00004424-199008000-00004

Kurtz, K. J., & Gentner, D. (2013). Detecting anomalous features in complex stimuli: The role of structured comparison. *Journal of Experimental Psychology: Applied*, *19*(3), 219–232. https://doi.org/10.1037/a0034395

Lenth, R. (2020). *emmeans: Estimated marginal means, aka least-squares means* (R package version 1.4.5).

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

Levett-Jones, T., Hoffman, K., Dempsey, J., Jeong, S. Y. S., Noble, D., Norton, C. A., Roche, J., & Hickey, N. (2010). The "five rights" of clinical reasoning: An educational model to enhance nursing students' ability to identify and manage clinically "at risk" patients. *Nurse Education Today*, *30*(6), 515–520. https://doi.org/10.1016/j.nedt.2009.10.020

Litchfield, D., & Ball, L. J. (2011). Rapid communication using another's gaze as an explicit aid to insight problem solving. *Quarterly Journal of Experimental Psychology*, *64*(4), 649–656. https://doi.org/10.1080/17470218.2011.558628

Litchfield, D., Ball, L. J., Donovan, T., Manning, D. J., & Crawford, T. (2010). Viewing another person's eye movements improves identification of pulmonary nodules in chest x-ray inspection. *Journal of Experimental Psychology: Applied*, *16*(3), 251–262. https://doi.org/10.1037/a0020082

Litchfield, D., & Donovan, T. (2016). Worth a quick look? Initial scene previews can guide eye movements as a function of domain-specific expertise but can also have unforeseen costs. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(7), 982–994. https://doi.org/10.1037/xhp0000202

Manning, D. J., Ethell, S. C., & Donovan, T. (2004). Detection or decision errors? Missed lung cancer from the posteroanterior chest radiograph. *British Journal of Radiology*, *77*(915), 231–235. https://doi.org/10.1259/bjr/28883951

Manning, D. J., 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. https://doi.org/10.1016/j.radi.2005.02.003

Mason, L., Pluchino, P., & Tornatora, M. C. (2015). Eye-movement modeling of integrative reading of an illustrated text: Effects on processing and learning. *Contemporary Educational Psychology*, *41*, 172–187. https://doi.org/10.1016/j.cedpsych.2015.01.004

Mason, L., Scheiter, K., & Tornatora, M. C. (2017). Using eye movements to model the sequence of text–picture processing for multimedia comprehension. *Journal of Computer Assisted Learning*, *33*(5), 443–460. https://doi.org/10.1111/jcal.12191

Monteiro, S. D., Sherbino, J., Schmidt, H., Mamede, S., Ilgen, J., & Norman, G. (2020). It's the destination: diagnostic accuracy and reasoning. *Advances in Health Sciences Education*, *25*, 19–29. https://doi.org/10.1007/s10459-019-09903-7

Myles-Worsley, M., Johnston, W. A., & Simons, M. A. (1988). The influence of expertise on X-ray image processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(3), 553–557. https://doi.org/10.1037/0278-7393.14.3.553

Nodine, C. F., & Kundel, H. L. (1987). Using eye movements to study visual search and to improve tumor detection. *RadioGraphics*, *7*(6), 1241–1250. https://doi.org/10.1148/radiographics.7.6.3423330

Nodine, C. F., Kundel, H. L., Mello-Thoms, C., Weinstein, S. P., Orel, S. G., Sullivan, D. C., & Conant, E. F. (1999). How experience and training influence mammography expertise. *Academic Radiology*, *6*(10), 575–585. https://doi.org/10.1016/S1076-6332(99)80252-9

Nodine, C. F., & Mello-Thoms, C. (2000). The nature of expertise in radiology. In J. Beutel, H. L. Kundel, & R. L. Van Metter (Eds.), *Handbook of Medical Imaging. Volume I. Physics and Psychophysics* (pp. 859–894). SPIE-The International Society for Optical Engineering.

Norman, G. (2005). Research in clinical reasoning: Past history and current trends. *Medical Education*, *39*(4), 418–427. https://doi.org/10.1111/j.1365-2929.2005.02127.x

Norman, G., Eva, K., Brooks, L., & Hamstra, S. (2006). Expertise in Medicine and Surgery. In K. A. Ericsson, N. Charness, P. J. Feltovich, & R. R. Hoffman (Eds.), *The Cambridge Handbook of Expertise and Expert Performance* (5th ed., pp. 339–353). Cambridge University Press.

Pasler, F. A. (1991). *Farbatlanten der Zahnmedizin. Band 5 Radiologie* (K. H. Rateitschak (ed.)). Georg Thieme Verlag Stuttgart.

Pasler, F. A., & Visser, H. (2007). *Pocket Atlas of Dental Radiology* (1st ed.). Georg Thieme Verlag Stuttgart.

Pinto, A., & Brunese, L. (2010). Spectrum of diagnostic errors in radiology. *World Journal of Radiology*, *2*(10), 377–383. https://doi.org/10.4329/wjr.v2.i10.377

Ramani, G. B., Daubert, E. N., Lin, G. C., Kamarsu, S., Wodzinski, A., & Jaeggi, S. M. (2020). Racing dragons and remembering aliens: Benefits of playing number and working memory games on kindergartners' numerical knowledge. *Developmental Science*, *23*(4), 1–17. https://doi.org/10.1111/desc.12908

Richter, J., Scheiter, K., Eder, T. F., Huettig, F., & Keutel, C. (2020). How massed practice improves visual expertise in reading panoramic radiographs in dental students: An eye tracking study. *PLoS ONE*, *15*(12), Article e0243060. https://doi.org/10.1371/journal.pone.0243060

Richter, J., Scheiter, K., & Eitel, A. (2016). Signaling text-picture relations in multimedia learning: A comprehensive meta-analysis. *Educational Research Review*, *17*, 19–36. https://doi.org/10.1016/j.edurev.2015.12.003

Roedinger, H. L. I., & Karpicke, J. D. (2006). The power of testing: Basic research and implications for educational practice. *Perspectives and Psychological Science*, *1*(3), 181–210. https://doi.org/10.1111/j.1745-6916.2006.00012.x

Ryan, J. T., Haygood, T. M., Yamal, J. M., Evanoff, M., O'Sullivan, P., McEntee, M., & Brennan, P. C. (2011). The "memory effect" for repeated radiologic observations. *American Journal of Roentgenology*, *197*(6), 985–991. https://doi.org/10.2214/AJR.10.5859

Scheiter, K., & Eitel, A. (2015). Signals foster multimedia learning by supporting integration of highlighted text and diagram elements. *Learning and Instruction*, *36*, 11–26. https://doi.org/10.1016/j.learninstruc.2014.11.002

Scheiter, K., & Eitel, A. (2016). Lernen mit Texten und Bildern. *Psychologische Rundschau*, *67*(2), 87–93. https://doi.org/10.1026/0033-3042/a000300

Scheiter, K., Schubert, C., & Schüler, A. (2018). Self-regulated learning from illustrated text: Eye movement modelling to support use and regulation of cognitive processes during learning from multimedia. *British Journal of Educational Psychology*, *88*(1), 80–94. https://doi.org/10.1111/bjep.12175

Schmidt, H. G., & Mamede, S. (2015). How to improve the teaching of clinical reasoning: A narrative review and a proposal. *Medical Education*, *49*(10), 961–973. https://doi.org/10.1111/medu.12775

Schmidt, H. G., & Mamede, S. (2020). How cognitive psychology changed the face of medical education research. *Advances in Health Sciences Education*, *25*(5), 1025–1043. https://doi.org/10.1007/s10459-020-10011-0

Schunk, D. H. (1987). Peer models and children's behavioral change. *Review of Educational Research*, *57*(2), 149–174. https://doi.org/10.3102/00346543057002149

Schunk, D. H., & Hanson, A. R. (1985). Peer models: Influence on children's self-efficacy and achievement. *Journal of Educational Psychology*, *77*(3), 313–322. https://doi.org/10.1037/0022-0663.77.3.313

Schuwirth, L. W. T., Durning, S. J., & King, S. M. (2020). Assessment of clinical reasoning: three evolutions of thought. *Diagnosis*, *7*(3), 191–196. https://doi.org/10.1515/dx-2019-0096

Seppänen, M., & Gegenfurtner, A. (2012). Seeing through a teacher's eyes improves students' imaging interpretation Marko. *Medical Education*, *46*(11), 1113–1114. https://doi.org/10.1111/medu.12041

Servant-Miklos, V. F. C. (2019). Problem solving skills versus knowledge acquisition: the historical dispute that split problem-based learning into two camps. *Advances in Health Sciences Education*, *24*(3), 619–635. https://doi.org/10.1007/s10459-018-9835-0

Sheridan, H., & Reingold, E. M. (2017). The holistic processing account of visual expertise in medical image perception: A review. *Frontiers in Psychology*, *8*, Article 1620. https://doi.org/10.3389/fpsyg.2017.01620

Smith, S., O'Tuathaigh, C., & Henn, P. (2017). Behind your very eyes: a response to Kok and Jarodzka. *Medical Education*, *51*(11), 1189–1189. https://doi.org/10.1111/medu.13341

Somppi, S., Törnqvist, H., Hänninen, L., Krause, C., & Vainio, O. (2012). Dogs do look at images: Eye tracking in canine cognition research. *Animal Cognition*, *15*(2), 163–174. https://doi.org/10.1007/s10071-011-0442-1

Stheeman, S. E., Mileman, P. A., van't Hot, M., & van der Stelt, P. F. (1996). Room for improvement? *Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology, and Endodontology*, *81*(2), 251–254. https://doi.org/10.1016/S1079-2104(96)80425-2

Swensson, R. G., Hessel, S. J., & Herman, P. G. (1977). Omissions in radiology: Faulty search or stingent reporting criteria? *Radiology*, *123*(3), 563–567. https://doi.org/10.1148/123.3.563

Swensson, R. G., Hessel, S. J., & Herman, P. G. (1985). The value of searching films without specific preconceptions. *Investigative Radiology*, *20*(1), 100–107. https://doi.org/10.1097/00004424-198501000-00024

Tamura, T., Inui, M., Nakase, M., Nakamura, S., Okumura, K., & Tagawa, T. (2005). Clinicostatistical study of carotid calcification on panoramic radiographs. *Oral Diseases*, *11*(5), 314–317. https://doi.org/10.1111/j.1601-0825.2005.01125.x

Turgeon, D. P., & Lam, E. W. N. (2016). Influence of experience and training on dental students' examination performance regarding panoramic images. *Journal of Dental Education*, *80*(2), 156–164. https://doi.org/10.1002/j.0022-0337.2016.80.2.tb06071.x

Vallo, J., Suominen-Taipale, L., & Huumonen, S. (2010). Prevalence of mucosal abnormalities of the maxillary sinus and their relationship to dental disease in panoramic radiography: results from the Health 2000 Health Examination Survey. *Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology, and Endodontology*, *109*(3), 80–87. https://doi.org/10.1016/j.tripleo.2009.10.031

van der Gijp, A., Ravesloot, C. J., Jarodzka, H., van der Schaaf, M. F., van der Schaaf, I. C., van Schaik, J. P. J., & ten Cate, T. J. (2017). How visual search relates to visual diagnostic performance: a narrative systematic review of eye-tracking research in radiology. *Advances in Health Sciences Education*, *22*, 765–787. https://doi.org/10.1007/s10459-016-9698-1

van der Gijp, A., van der Schaaf, M. F., van der Schaaf, I. C., Huige, J. C. B. M., Ravesloot, C. J., van Schaik, J. P. J., & ten Cate, T. J. (2014). Interpretation of radiological images: towards a framework of knowledge and skills. *Advances in Health Sciences Education*, *19*, 565–580. https://doi.org/10.1007/s10459-013-9488-y

van Geel, K., Kok, E. M., Dijkstra, J., Robben, S. G. F., & van Merriënboer, J. J. G. (2017). Teaching systematic viewing to final-year medical students improves systematicity but not coverage or detection of radiologic abnormalities. *Journal of the American College of Radiology*, *14*(2), 235–241. https://doi.org/10.1016/j.jacr.2016.10.001

van Gog, T. (2014). The signaling (or cueing) principle in multimedia learning. In R. E. Mayer (Ed.), *The cambridge handbook of multimedial learning* (2nd ed., pp. 263–278). Cambridge University Press.

van Gog, T., Jarodzka, H., Scheiter, K., Gerjets, P., & Paas, F. (2009). Attention guidance during example study via the model's eye movements. *Computers in Human Behavior*, *25*(3), 785–791. https://doi.org/10.1016/j.chb.2009.02.007

van Gog, T., & Rummel, N. (2010). Example-based learning: Integrating cognitive and social-cognitive research perspectives. *Educational Psychology Review*, *22*(2), 155–174. https://doi.org/10.1007/s10648-010-9134-7

van Marlen, T., van Wermeskerken, M., Jarodzka, H., & van Gog, T. (2016). Showing a model's eye movements in examples does not improve learning of problem-solving tasks. *Computers in Human Behavior*, *65*, 448–459. https://doi.org/10.1016/j.chb.2016.08.041

van Marlen, T., van Wermeskerken, M., Jarodzka, H., & Van Gog, T. (2018). Effectiveness of eye movement modeling examples in problem solving: The role of verbal ambiguity and prior knowledge. *Learning and Instruction*, *58*, 274–283. https://doi.org/10.1016/j.learninstruc.2018.07.005

van Merriënboer, J. J. G., & Kirschner, P. A. (2007). *Ten steps to complex learning. A systematic approach to four-component instructional design.* (L. Akers (ed.)). Lawrence Erlbaum Associates.

Vock, P., & Woermann, U. (2016). *RadioSurf*.

Waite, S., Farooq, Z., Grigorian, A., Sistrom, C., Kolla, S., Mancuso, A., Martinez-Conde, S., Alexander, R. G., Kantor, A., & Macknik, S. L. (2020). A review of perceptual expertise in radiology-how it develops, how we can test it, and why humans still matter in the era of artificial intelligence. *Academic Radiology*, *27*(1), 26–38. https://doi.org/10.1016/j.acra.2019.08.018

Waite, S., Grigorian, A., Alexander, R. G., Macknik, S. L., Carrasco, M., Heeger, D. J., & Martinez-Conde, S. (2019). Analysis of perceptual expertise in radiology – Current knowledge and a new perspective. *Frontiers in Human Neuroscience*, *13*, Article 213. https://doi.org/10.3389/fnhum.2019.00213

Wedel, M., Yan, J., Siegel, E. L., & Li, H. A. (2016). Nodule detection with eye movements. *Journal of Behavioral Decision Making*, *29*(2–3), 254–270. https://doi.org/10.1002/bdm.1935

Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta Psychologica*, *41*, 67–85. https://doi.org/10.1016/0001-6918(77)90012-9

Williams, I. J. (2013). Appendicular skeleton : ABCs image interpretation search strategy. *South African Radiographer*, *51*(2), 9–14.

Wolfe, J. M. (2012). Saved by a log: How do humans perform hybrid visual and memory search? *Psychological Science*, *23*(7), 698–703. https://doi.org/10.1177/0956797612443968

Wolfe, J. M. (2016). Use-inspired basic research in medical image perception. *Cognitive Research: Principles and Implications*, *1*(17), 1–9. https://doi.org/10.1186/s41235-016-0019-2

Wolfe, J. M., Evans, K. K., Drew, T., Aizenman, A., & Josephs, E. (2016). How do radiologists use the human search engine? *Radiation Protection Dosimetry*, *169*(1), 24–31. https://doi.org/10.1093/rpd/ncv501

Wolfe, J. M., Horowitz, T. S., & Kenner, N. M. (2005). Rare items often missed in visual searches. *Nature*, *435*(7041), 439–440. https://doi.org/10.1038/435439a

Wu, C. C., & Wolfe, J. M. (2019). Eye movements in medical image perception: A selective review of past, present and future. *Vision*, *3*(2), 1–15. https://doi.org/10.3390/vision3020032

12. Appendices

A Parameters and their explanation of the model for Hypotheses 1, 2a, 2b, 2c & 3

Parameters	Explanation of parameters
y_{ijkl}	Estimated value for each student i, at time j in the specific group k, for the location l while considering the conceptual knowledge of student i
β_0	Intercept across students for the reference categories (pre-test, control group, central anomalies)
β_1	Fixed effect of time (pre- vs. post-test)
β_2	Fixed effect of group (control vs. intervention group)
β_3	Fixed effect of location (central vs. peripheral anomalies)
β_4	Interaction of time and group
β_5	Interaction of time and location
β_6	Interaction of group and location
β_7	Three-way interaction of time, group and location
β_8	Conceptual knowledge as covariate to control for different levels
v_{0i}	Random effect: Individual intercept for each student
v_{1i}	Random effect: Individual slope over time for each student
ε_{ijkl}	Error term

B Means and standard deviations of gaze measures for log-transformed values

		Control group				Intervention group			
		Pre-test		Post-test		Pre-test		Post-test	
		central	peripheral	central	peripheral	central	peripheral	central	peripheral
Number of	Mean	0.84	2.21	1.17	2.22	1.02	2.17	0.91	2.07
fixations	SD	0.26	0.18	0.19	0.16	0.22	0.25	0.25	0.23
Fixation time	Mean	7.28	8.37	7.58	8.37	7.43	8.31	7.42	8.28
(ms)	SD	0.48	0.22	0.35	0.24	0.27	0.29	0.29	0.22

C Model parameters from the linear mixed models of coverage and gaze measures

		Gaze coverage rate		Number of fixations		Fixation time		Time to first fixation	
		Estimate (SE)	t value (df)	Estimate (SE)	t value (df)	Estimate (SE)	t value (df)	Estimate (SE)	t value (df)
Fixed effects	Intercept	43.54 (2.25)	t(83.81) = 19.35***	0.89 (0.09)	t(109.11) = 9.42***	7.26 (0.12)	t(105.85) = 58.56***	25987.05 (2459.31)	t(122.99) = 10.57***
	Time[1]	-2.85 (1.41)	t(61.47) = -2.02*	0.34 (0.07)	t(127.78) = 4.88***	0.29 (0.10)	t(118.13) = 3.04**	-1893.38 (1943.07)	t(154.67) = -0.97
	Group[2]	-0.85 (1.31)	t(47.81) = -.65	0.17 (0.07)	t(91.12) = 2.47*	0.15 (0.09)	t(86.91) = 1.63	896.21 (1864.53)	t(106.35) = 0.48
	Location[3]			1.36 (0.06)	t(94.00) = 21.40***	1.09 (0.08)	t(94.00) = 13.02***	-15300.32 (1846.38)	t(140.82) = -8.29***
	Conceptual knowledge	0.45 (0.18)	t(86.71) = 2.53*	0.00 (0.01)	t(86.50) = -.062	0.00 (0.01)	t(80.02) = 0.22	265.89 (176.59)	t(106.08) = 1.51
	Time x Group	4.08 (1.63)	t(51.92) = 2.50*	-0.45 (0.08)	t(122.69) = -5.36***	-0.31 (0.12)	t(112.66) = -2.66**	5441.30 (2346.03)	t(146.71) = 2.32*
	Time x Location			-0.32 (0.09)	t(94.00) = -3.54***	-0.30 (0.12)	t(94.00) = -2.57*	3279.71 (2611.17)	t(140.82) = 1.26
	Group x Location			-0.22 (0.08)	t(94.00) = -2.73**	-0.21 (0.10)	t(94.00) = -2.02*	1593.15 (2284.78)	t(140.82) = 0.70

	Time x Group x Location		0.33 (0.11)	t(94.00) = 3.01**	0.28 (0.15)	t(94.00) = 1.91	-6915.16 (3231.16)	t(140.82) = -2.14*
Random effects	Individual intercept (SD)	4.00	0.14		0.19		3001.87	
	Individual slope over time (SD)	4.77	0.08		0.16		690.32	
	Correlation of intercept and slope	-0.35	-0.67		-0.77		-1.00	
	Residual variance (SD)	4.83	0.19		0.24		5383.07	
	Individual intercept for OPTs (SD)	1.40						

[1]Dummy-coded (pre-test as reference category); [2]Dummy-coded (control group as reference category); [3]Dummy-coded (central anomalies as reference category); *p < .05, **p < .01, ***p < .001

D Model parameters from the generalized linear mixed models of diagnostic performance

		Detection rate		Number of false positive markings	
		Estimate (SE)	z value	Estimate (SE)	z value
Fixed effects	Intercept	-0.25 (0.18)	-1.37	2.42 (0.22)	10.78***
	Time[1]	-0.14 (0.10)	-1.42	0.32 (0.13)	2.42*
	Group[2]	0.05 (0.09)	0.55	-0.13 (0.13)	-0.99
	Location[3]	-0.18 (0.11)	-1.60	-0.91 (0.12)	-7.88***
	Conceptual knowledge	0.03 (0.01)	1.85	-0.01 (0.02)	-0.29
	Time x Group	0.26 (0.12)	2.12*	0.47 (0.16)	2.95**
	Time x Location	0.23 (0.16)	1.47	0.20 (0.15)	1.39
	Group x Location	0.28 (0.14)	1.93	-0.14 (0.15)	-0.90
	Time x Group x Location	-0.23 (0.20)	-1.15	0.11 (0.19)	0.61
Random effects	Individual intercept (SD)	0.25		0.38	
	Individual slope over time (SD)	0.31		0.43	
	Correlation of intercept and slope	0.10		-0.21	

[1]Dummy-coded (pre-test as reference category); [2]Dummy-coded (control group as reference category); [3]Dummy-coded (central anomalies as reference category); *p < .05, **p < .01, ***p < .001