Chest X-ray evaluation training: impact of normal and abnormal image ratio and instructional sequence

Koos van Geel,[1] Ellen M Kok,[2] Abdullah D Aldekhayel,[1] Simon G F Robben[1] & Jeroen J G van Merriënboer[3]

CONTEXT Medical image perception training generally focuses on abnormalities, whereas normal images are more prevalent in medical practice. Furthermore, instructional sequences that let students practice prior to expert instruction (inductive) may lead to improved performance compared with methods that give students expert instruction before practice (deductive). This study investigates the effects of the proportion of normal images and practice–instruction order on learning to interpret medical images. It is hypothesised that manipulation of the proportion of normal images will lead to a sensitivity–specificity trade-off and that students in practice-first (inductive) conditons need more time per practice case but will correctly identify more test cases.

METHODS Third-year medical students (n = 103) learned radiograph interpretation by practising cases with, respectively, 30% or 70% normal radiographs prior to expert instruction (practice-first order) or after expert instruction (instruction-first order). After training, students performed a test (60% normal) and sensitivity (% of correctly identified abnormal radiographs), specificity (% of correctly identified normal radiographs), diagnostic performance (% of correct diagnoses) and case duration were measured.

RESULTS The conditions with 30% of normal images scored higher on sensitivity but the conditions with 70% of normal images scored higher on specificity, indicating a sensitivity and specificity trade-off. Those who participated in inductive conditions took less time per practice case but more per test case. They had similar test sensitivity, but scored lower on test specificity.

CONCLUSIONS The proportion of normal images impacted the sensitivity–specificity trade-off. This trade-off should be an important consideration for the alignment of training with future practice. Furthermore, the deductive conditions unexpectedly scored higher on specificity when participants took less time per case. An inductive approach did not lead to higher diagnostic performance, possibly because participants might already have relevant prior knowledge. Deductive approaches are therefore advised for the training of advanced learners.

Medical Education 2019: 53: 153–164
doi: 10.1111/medu.13756

[1]Department of Radiology, Maastricht University Medical Center, Maastricht, the Netherlands
[2]Department of Education, Utrecht University, Utrecht, the Netherlands
[3]School of Health Professions Education, Department of Educational Research and Development, Maastricht University, Maastricht, the Netherlands

Correspondence: Koos van Geel, Department of Radiology, Maastricht University Medical Centre, PO Box 5800, 6202 AZ Maastricht, the Netherlands. Tel: +00 31 43 388 5776; E-mail: k.vangeel@maastrichtuniversity.nl.

INTRODUCTION

The interpretation of medical images, such as electrocardiograms, pathology slices or radiographs, is an important part of everyday medical practice.[1–3] Research on medical image interpretation has primarily focused on characteristics of visual expertise.[3,4] In such research, novices and experts in image interpretation are compared and the experts' performance is superior to that of novices. Experts also show more efficient viewing behaviour.[5] Although such research on visual expertise provides invaluable information on how *learning* to interpret images takes place, it does not provide straightforward answers to questions regarding *teaching* medical image perception. The current study aims to add to the literature regarding: (i) the 'what' content of medical image perception training; and (ii) the 'how' instructional design of medical image perception training.

The content of medical image perception training

Concerning the content of medical image perception training, there is generally a large emphasis on abnormal images.[2] Only a small amount of time in medical curricula is devoted to teaching image interpretation,[6] whereas a vast amount of anatomy and (patho)physiology needs to be covered. Although it might be time efficient, this emphasis on abnormal images may also give students a wrong impression about the prevalence of diseases in medical practice. In reality, many images in everyday clinical practice on a ward or in an emergency department are found to be normal or do not contain significant or relevant pathology.[7–9] This mismatch between low prevalence of diseases in clinical practice and the emphasis on abnormal images during training can impact students' performance in practice. Indeed, Pusic et al.[2] have shown that a change in the proportion of abnormal practice cases alters the sensitivity (proportion of correctly identified abnormal images out of the total number of abnormal images) and specificity (proportion of correctly identified normal images out of the total number of normal images) of the performance of emergency residents. The residents who practised with predominantly abnormal images had higher sensitivity, whereas the residents who practised with predominantly normal images had higher specificity. The emergency residents in the study by Pusic et al.[2] already had some experience in interpreting medical images and might have learned about the low prevalence of diseases in clinical practice. It is not yet known to what extent medical students are impacted by the proportion of normal images in training. It is expected that the performance of more novice students potentially increases even more when they are trained with a high proportion of normal images in medical image perception training.

Instructional design of medical image perception training

The instructional design of medical image perception training, like most educational experiences, often consists of a presentation by an expert, practise of the task by learners and feedback. *When* to provide expert instruction and practice for an effective educational experience remains a debate in medical education.[10] Direct or deductive-expository instruction, which starts with the expert instruction followed by a practice phase, is advocated for more advanced learners, when instructional time is limited and when a deep level of understanding is not strictly necessary.[11]

By contrast, inductive approaches such as problem-based learning and guided discovery learning[12] offer practice prior to instruction. As students first practise, they will have to figure out solutions for themselves instead of only implementing a solution presented by an expert. Students may fail to find the solution and will need more time to complete a practice case. However, this failure may be considered productive.[13] Students are fully immersed in the problem when searching for the solution. This productive failure can therefore lead to a deeper understanding and long-term retention of knowledge.[14] The benefits of productive failure indeed have been shown in research in mathematics education.[15] Despite the theoretical benefit of inductive approaches, most medical image perception training still uses deductive approaches. It is therefore not known if productive failure can be induced in medical students who are learning to interpret medical images.

The present study

In this study, the effects of the proportion of normal images (30% versus 70% normal) and instructional sequence (deductive versus inductive), in a chest radiograph perception training, on the performance of third-year medical students were investigated.

Research questions

1 What are the effects of the proportion of normal images in a practice phase of medical image perception training on third-year medical students' performance?
2 What are the effects of instructional sequencing (inductive or deductive) in medical image perception training on third-year medical students' performance?

The students' performance was defined as sensitivity, specificity, diagnostic performance and case duration on a subsequent test.

Hypotheses

In line with Pusic et al.,[2] we hypothesise that:

1 Students practising with a low proportion of normal images will have higher sensitivity scores, whereas students practising with a high proportion of normal images will have higher specificity scores.
2 Students in inductive conditions will have higher sensitivity, specificity and diagnostic performance than students in deductive conditions.

Concerning students' performance during the practice phase, students in the inductive conditions will be engaged in the act of productive failure, we hypothesise that this should result in:

1 Lower sensitivity, specificity and diagnostic performance in the inductive conditions.
2 Image interpretation during the practice phase will take more time.
3 Students will need more time per case on those they misinterpret compared with cases they correctly interpret, which will reflect productive failure. This difference will be higher for students in inductive conditions.

We do not hypothesise that any interaction effects will occur in our analyses, and these terms are therefore exploratory.

METHOD

This 2 × 2 design tested the effects of proportion of normal images (practising with a proportion of 30% normal images [condition 1] versus a proportion of 70% normal images [condition 2]) and instructional sequence (a practice-first [inductive] versus an instruction-first [deductive] sequence [conditions 3 and 4]) (Fig. 1). After the training, students' sensitivity, specificity, diagnostic performance and case duration were measured in a test with a proportion of normal images that is typical of everyday clinical practice.

Participants

A total of 103 third-year medical students took part in this study (69% female; mean age = 22.5 ± 2.43 years) from Maastricht University in the Netherlands. All students were approached via announcements prior to regular lectures and via announcements on the electronic learning environment of Maastricht University in September 2016. None of the participants had yet received any formal training in interpreting chest radiographs. Participants were randomly assigned to the four experimental conditions in a 2 × 2 design (Fig. 1).

Two of the conditions started with the practice phase, consisting of practising with 20 chest radiographs. The proportion of normal radiographs during the practice phase was manipulated (70% normal radiographs versus 30% normal radiographs). The other two conditions started with the instruction phase, consisting of a video lecture, and subsequently practised with a set of either 70% normal or 30% normal images, yielding a full 2 × 2 design. Participants received a €20 gift voucher after the experiment as compensation. All participants signed an informed consent, and the study was approved by the Ethical Review Board of the Dutch Association for Medical Education (NVMO-ERB), file number 763.

Materials

Video lecture

During the instruction phase, a video lecture was used. This video lecture was designed for this experiment by AA and SGFR. The video covered the basics of chest radiograph interpretation and the radiologic manifestations of eight common abnormalities: pneumonia, pneumothorax, pleural effusion, atelectasis, lung tumours, cardiomegaly, emphysema and bilateral hilar lymphadenopathy. Two normal chest radiographs and two examples of each abnormality were used in the video, which totalled 18 radiographs. The video had a duration of 23 minutes and participants saw the video only once. Participants were not allowed to stop, rewind or fastforward the video. Furthermore, participants were

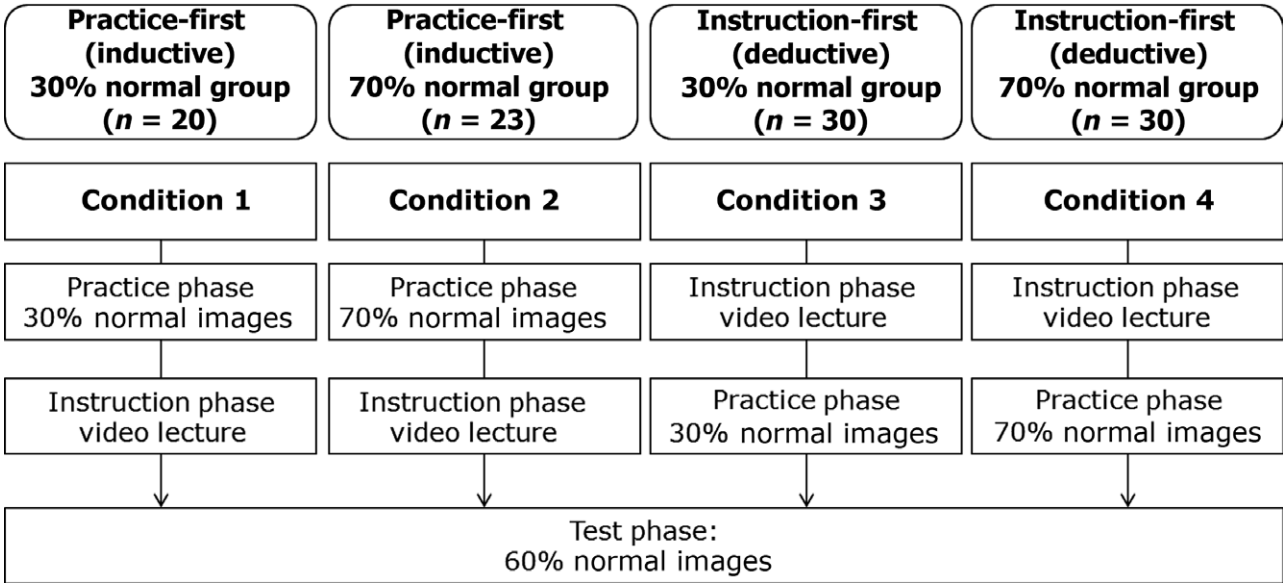

Figure 1 Flowchart of the 2 × 2 design on the four experimental conditions

not allowed to make notes. The video lecture was shown individually to participants using WINDOWS Movie Player 12 (Microsoft Corp., Redmond, WA, USA).

Radiological images

The radiographs used in this experiment originated from a teaching file consisting from over 400 chest radiographs from the radiology department at the Maastricht University Medical Center. All radiographs were stripped of any patient information and were selected by KvG and SGFR. Radiographs were selected to have no abnormalities (i.e. normal images) or only one type of the eight previously mentioned cardiopulmonary pathologies. The cases have been used in previous investigations involving third-year medical students, as well as final-year medical students.[16,17] In the investigation with final-year medical students, learning effects were visible after practising with 10 cases and a video lecture. To ensure a learning effect in third-year medical students, the number of cases in the practice phase was doubled. The images used in this investigation are available on request from the first author (KvG).

In the practice phase, participants interpreted 20 chest radiographs; 12 chest radiographs were identical for each of the four conditions with half of these identical chest radiographs being normal images. In conditions 1 and 3, for the 30% normal images, the other eight chest radiographs were abnormal. In conditions 2 and 4, for the 70% normal images, the other eight chest radiographs were normal. An example of a chest X-ray used in the experiment can be found online in the Supporting Information (Figure S1).

The test phase consisted of 20 chest radiographs, of which 60% were normal images. In daily practice, normal images predominate over abnormal images.[7] The order of the 20 chest radiographs was randomised per participant. The cases were heterogeneous in variance: abnormal cases, $F(7, 102) = 13.4$, $p < 0.001$; normal cases, $F(11, 102) = 35.4$, $p < 0.001$; diagnostic performance, $F(7, 102) = 25.3$, $p < 0.001$. Calculation of Cronbach's alpha would produce unreliable estimates.[18] Instead, Macdonalds' Ω_t was calculated, which can be interpreted similarly to a Cronbach's alpha. The Ω_t of abnormal cases was 0.63, the Ω_t of normal cases was 0.67 and the Ω_t of diagnostic performance was 0.41. A table with the characteristics of the test phase (discrimination of the cases, mean percentage correctly identified, average case duration and diagnosis per case) can be found online as Supporting Information (Table S1).

Measures

Sensitivity and specificity

Sensitivity in the practice phase and test phase was defined as the proportion of abnormal radiographs correctly identified as abnormal. Specificity in the practice phase and test phase was defined as the proportion of normal radiographs correctly identified as normal.

Diagnostic performance

If the participants deemed a radiograph to be abnormal during the practice phase or test phase, they were requested to type their most probable diagnosis via a free text form. A coding scheme for correct diagnoses and their respective synonyms was developed by KvG and SGFR. All correctly diagnosed radiographs were subsequently coded as 1, all incorrect answers were coded as 0. To calculate the diagnostic performance of participants, all diagnosis scores were summed and divided by the total number of eight abnormal radiographs. For the diagnosis scores of the practice phase, only the six abnormal cases that were identical in all four conditions were used.

Average case duration

The time needed by participants to interpret a chest radiograph and provide answers was registered and averaged for the 12 (normal and abnormal) identical radiographs (cases) in the practice phase and the 20 radiographs (cases) in the test phase.

Procedure

The experiment was conducted in 11 sessions with a maximum of 10 students per experimental session. Every participant worked on a desktop computer with a 22″ LCD (liquid crystal display) screen with a resolution of 1650 × 1080 pixels using the Qualtrics software (Qualtrics, Provo, Utah, USA).[19] Each session started with a short briefing of 5 minutes in which the procedure was delineated, and participants subsequently provided written consent. Participants were not informed about the proportion of normal images in the practice phase or test phase. The order of the cases in the test phase was randomised per participant by the QUALTRICS software. Participants worked individually throughout the whole experiment.

During the practice phase, participants had a maximum of 80 seconds to interpret each of the 20 chest radiographs and to report if they were normal or abnormal images. If the image was abnormal, they were required to report the most probable diagnosis. The time limit of 80 seconds was based on a previous investigation with third-year medical students who took an average of 52.6 seconds (standard deviation [SD] 20.6) to interpret a case.[16] Based on these numbers, the probability of not completing a case within 80 seconds would be 0.09, which was considered acceptable. After 80 seconds, a new page was automatically loaded that informed participants whether the radiograph was normal or abnormal. Furthermore, if the radiograph was abnormal, the diagnosis was given. Participants had a maximum of 10 seconds to read the feedback page. After 10 seconds, the feedback page closed automatically and the next radiograph was loaded. In the instruction phase, participants individually watched the video lecture with 18 example chest radiographs. When participants had completed both the practice phase and the instruction phase, they had a short break of 5 minutes.

After the break, participants entered the test phase, in which participants had a maximum of 90 seconds to interpret and report every radiograph. After 90 seconds, the next case was automatically loaded. Participants did not receive any feedback about interpreted images during the test phase.

Analyses

For the analyses, 2 × 2 analyses of variance (ANOVAs) were performed with the factors instructional sequence (to practice-first order (inductive) and instruction-first order (deductive)) and proportion (a proportion of 30% normal images versus a proportion of 70% normal images) on the outcome measures of the test phase and practice phase. The sensitivity scores of the test phase and practice phase was negatively skewed; the lowest $Z_{skewness}$ score for the test phase sensitivity was found in the instruction-first (deductive), condition 4, with 70% normal images and was −4.37. The lowest $Z_{skewness}$ score for the practice phase sensitivity was found in the instruction-first (deductive), condition 3 with 30% normal images and was −2.40. As there is currently no reasonable non-parametric alternative to a 2 × 2 ANOVA and that ANOVA analyses are generally robust for skewness, these skewness levels were tolerated. As a measure of effect size an η_p^2 was used, with 0.01 indicating a small effect, 0.06 indicating a moderate effect and 0.14 indicating a large effect.[19,20]

To analyse differences between the four conditions in case durations for cases divided into correctly identified versus incorrectly identified cases, a full-factorial binary logistic regression analysis of the practice phase was performed, with discrimination score (correct versus incorrect) as the dependent variable and instructional sequence, proportion and case duration as independent variables.

RESULTS

Results of the test phase

The descriptors and the results of the 2 × 2 ANOVA per test-phase measure can be found in Table 1. Furthermore, the descriptors of the test-phase are visualised as violin plots to be found online as Supporting Information (Figure S2).

On *sensitivity*, a main effect of proportion of normal images was found, in favour of practising with 30% normal images; the found main effect is in line with hypothesis 1. There was no main effect of sequence. No significant interaction effect between proportion and instructional sequence was found.

On *specificity*, a main effect of proportion of images was found in favour of practising with 70% normal images; the found main effect is in line with hypothesis 1. Furthermore, a main effect of sequence, now in favour of instruction-first (deductive), conditions 3 and 4, was found. No significant interaction between proportion of normal images and sequence was found.

On *diagnostic performance*, no main effect of proportion was found, which contrasted with hypothesis 1. There was no main effect of sequence, which contrasted with hypothesis 2. No significant interaction effect between proportion and sequence was found.

On *average case duration*, no main effect of proportion was found. A significant main effect of sequence was found; the average case duration was higher in the practice-first (inductive), conditions 1 and 2. No significant interaction effect between proportion and sequence was found.

The time limit for interpreting cases was reached in five out of 400 cases for the group practice-first (inductive), condition 1 with 30% normal images, five out of 460 cases for the group practice-first (inductive), condition 2 with 70% normal images, eight out of 600 cases for the group instruction-first (deductive), condition 3 with 30% normal images, and seven out of 600 cases for the group instruction-first (deductive), condition 4 with 70% normal images. The number of cases in which the time limit was reached did not differ between the four conditions, χ^2 (3, n = 2060) = 0.15, p = 0.99.

Table 1 Descriptors and results of 2 × 2 ANOVA of test-phase measures

	Practice-first order (inductive)		Instruction-first order (deductive)				
	Condition 1 30% normal (n = 20)	Condition 2 70% normal (n = 23)	Condition 3 30% normal (n = 30)	Condition 4 70% normal (n = 30)	2 × 2 ANOVA		
Variable	Mean ± SD	Mean ± SD	Mean ± SD	Mean ± SD	Main effect of proportion of normal images	Main effect of instructional sequence	Interaction effect
Sensitivity (%)	97.5 ± 5.13	89.1 ± 9.46	96.7 ± 6.51	89.6 ± 8.73	F (1, 99) = 24.97, p < 0.001, η_p^2 = 0.20	F (1, 99) = 0.02, p = 0.90, η_p^2 < 0.001	F (1, 99) = 0.17, p = 0.68, η_p^2 < 0.001
Specificity (%)	52.5 ± 15.3	65.2 ± 14.8	58.3 ± 18.0	72.8 ± 13.3	F (1, 99) = 20.70, p < 0.001, η_p^2 = 0.17	F (1, 99) = 5.03, p = 0.03, η_p^2 = 0.05	F (1, 99) = 0.08, p = 0.77, η_p^2 < 0.001
Diagnostic performance (%)	53.8 ± 18.1	48.9 ± 15.0	52.9 ± 17.9	52.5 ± 13.2	F (1, 99) = 0.67, p = 0.42, η_p^2 < 0.001	F (1, 99) = 0.18, p = 0.67, η_p^2 < 0.001	F (1, 99) = 0.47, p = 0.49, η_p^2 < 0.001
Case duration (s)	35.1 ± 9.31	34.0 ± 10.7	30.4 ± 7.17	29.8 ± 8.25	F (1, 99) = 1.57, p = 0.21, η_p^2 < 0.001	F (1, 99) = 9.61, p = 0.003, η_p^2 = 0.09	F (1, 99) < 0.001, p = 0.95, η_p^2 < 0.001

SD, standard deviation.

Results of the practice phase

The descriptors and the results of the 2 × 2 ANOVA of the practice phase can be found in Table 2. Furthermore, the descriptors of the practice phase measures are visualised as violin plots to be found online as Supporting Information (Figure S3).

On *sensitivity*, no main effect of proportion was found. Furthermore, a main effect of instructional sequence, in favour of the instruction-first order, was found. This is in line with hypothesis 3. Finally, a marginally significant interaction between proportion and instructional sequence was found, in favour of the group instruction-first (deductive), condition 3 with 30% normal images.

On *specificity*, a significant main effect of proportion was found, in favour of practising with 70% normal images. By contrast with hypothesis 3, there was no main effect of instructional sequence. No interaction effect of proportion of normal images and instructional sequence was found.

On *diagnostic performance*, no main effect of proportion was found. A main effect of instructional sequence in favour of instruction-first (deductive), conditions 3 and 4, was found, in line with hypothesis 3. No significant interaction effect of proportion and instructional sequence was found.

On *case duration*, no main effect of proportion was found. Unexpectedly and by contrast with hypothesis 4, the participants in practice-first (inductive) conditions 1 and 2 took less time to complete the practice cases than the participants in instruction-first (deductive) conditions 3 and 4. A main effect of instructional sequence was found; the average case duration in the instruction-first (deductive), conditions 3 and 4 groups, was higher. Finally, a marginally significant interaction effect was found; the average case duration of the practice-first (inductive), condition 2 group with 70% normal images, was the lowest.

The number of cases in which the time limit was reached per condition can be found in Table 3. The time limit for *interpreting* cases was more often reached in the instruction-first (deductive), conditions 3 and 4 groups, χ^2 (1, n = 2060) = 8.3, p = 0.004. The number of cases in which the time limit for *reading the feedback* was reached did not

Table 2 Practice-phase descriptors and 2 × 2 ANOVA results for each separate condition

	Practice-first order (inductive)		Instruction-first order (deductive)				
	Condition 1 30% normal (n = 20)	Condition 2 70% normal (n = 23)	Condition 3 30% normal (n = 30)	Condition 4 70% normal (n = 30)	2 × 2 ANOVA		
Variable	Mean ± SD	Mean ± SD	Mean ± SD	Mean ± SD	Main effect of proportion of normal images	Main effect of instructional sequence	Interaction effect
Sensitivity (%)	80.0 ± 15.9	71.7 ± 19.7	88.9 ± 14.0	92.2 ± 10.0	F (1, 99) = 24.97, p < 0.001, η_p^2 = 0.20	F (1, 99) = 0.02, p = 0.90, η_p^2 < 0.001	F (1, 99) = 0.17, p = 0.68, η_p^2 < 0.001
Specificity (%)	36.7 ± 22.7	45.6 ± 17.6	35.0 ± 25.2	44.4 ± 24.5	F (1, 99) = 20.70, p < 0.001, η_p^2 = 0.17	F (1, 99) = 5.03, p = 0.03, η_p^2 = 0.05	F (1, 99) = 0.08, p = 0.77, η_p^2 < 0.001
Diagnostic performance (%)	36.7 ± 18.4	31.9 ± 13.2	57.2 ± 16.8	53.3 ± 18.2	F (1, 99) = 0.67, p = 0.42, η_p^2 < 0.001	F (1, 99) = 0.18, p = 0.67, η_p^2 < 0.001	F (1, 99) = 0.47, p = 0.49, η_p^2 < 0.001
Case duration (s)	44.5 ± 9.40	40.1 ± 9.10	47.4 ± 8.82	49.6 ± 7.90	F (1, 99) = 1.57, p = 0.21, η_p^2 < 0.001	F (1, 99) = 9.61, p = 0.003, η_p^2 = 0.09	F (1, 99) < 0.001, p = 0.95, η_p^2 < 0.001

SD, standard deviation.

Table 3 Occurrence of being not within time limits during the practice phase per condition

	Practice-first order (inductive)		Instruction-first order (deductive)	
Practice phase time limits reached	**Condition 1 30% normal (*n* = 460)**	**Condition 2 70% normal (*n* = 400)**	**Condition 3 30% normal (*n* = 600)**	**Condition 4 70% normal (*n* = 600)**
Interpreting cases	34	27	74	56
Reading the feedback	2	3	2	10

differ between the instructional sequences, χ^2 (1, n = 2060) = 1.1, p = 0.30.

Occurrence of productive failure during the practice phase

The average case durations for correct and incorrect interpretations of the 12 identical cases of the practice phase can be found in Table 4.

Furthermore, the results of the binary logistic regression can also be found in Table 4.

The binary logistic regression analysis showed that in both instruction-first (deductive), conditions 3 and 4, participants took longer to identify cases than in both practice-first (inductive), conditions 1 and 2. Furthermore, a main effect of case duration was found, indicating that correctly identified cases

Table 4 Results of the binary logistic regression with correctly identified cases as outcome variable.

	Correctly identified			
	No		Yes	
Conditions	***n***	**Mean ± SD**	***n***	**Mean ± SD**
1 Practice-first (inductive), 30% normal	114	44.0 (22.9)	139	38.0 (17.9)
2 Practice-first (inductive), 70% normal	100	51.1 (23.5)	120	40.2 (17.8)
3 Instruction-first (deductive), 30% normal	114	62.0 (18.9)	216	44.9 (19.2)
4 Instruction-first (deductive), 70% normal	137	58.0 (18.6)	193	42.2 (19.3)
Binary logistic regression analysis	**B (SE)**	**d.f.**	**p**	**OR (95% CI)**
Intercept	1.32 (0.34)	1	<0.001	3.73
Instructional sequence	1.08 (0.48)	1	0.025	2.93 (1.14–7.51)
Prevalence	−0.53 (0.40)	1	0.24	0.59 (0.25–1.41)
Case duration	−0.025 (0.0070)	1	<0.001	0.98 (0.96–0.99)
Instructional sequence * prevalence	1.16 (0.68)	1	0.089	3.19 (0.84–12.2)
Instructional sequence * case duration	−0.016 (0.0090)	1	0.081	0.98 (0.97–1.00)
Prevalence * case duration	0.01 (0.0090)	1	0.26	1.01 (0.99–1.03)
Instructional sequence * prevalence * case duration	−0.014 (0.013)	1	0.28	0.99 (0.96–1.01)

CI, confidence interval; d.f., degrees of freedom; OR, odds ratio; p probability; SD, standard deviation; SE, standard error.
* χ^2 (7) = 132.66, R^2 = 0.11 (Cox & Snell), 0.15 (Nagelkerke).

were interpreted faster than incorrectly identified cases.

By contrast with hypothesis 5, all two-way and three-way interaction terms were non-significant, indicating that the found main effects of instructional sequence and case duration were similar for all four conditions.

DISCUSSION

In this experiment, the proportion of normal images during a practice phase and the instructional sequence of medical image perception training were manipulated. The effect of changing the proportion of normal images, which was previously found by Pusic et al.[2] in a sample of residents, was replicated in our sample of medical students. In line with the hypothesis 1, sensitivity scores were highest in the conditions with a low proportion of normal images and specificity scores were highest in the conditions with a high proportion of normal images. It was thus found that students who train with more normal images are less likely to make false positive errors (reporting abnormalities that are not present), whereas students training with mostly abnormal images are less likely to miss abnormalities, a phenomenon known as a 'criterion shift'.[21,22] One of the first and most important steps in interpreting medical images is categorisation of the image into normal or abnormal.[23,24] For this categorisation, a decision criterion is used, which is influenced by previous experiences.[23] Medical image perception training is generally the first experience that students have of interpreting medical images. A mismatch between the prevalence of abnormalities in the training[2] and of medical images in everyday clinical practice[7–9] can easily result in students being trained with a suboptimal criterion. Our study shows that a short 20-item training session can already have an impact on this criterion.[25]

With regard to the effects of instructional sequences on performance measures, the deductive sequence conditions (3 and 4) led to higher student performance scores than the inductive sequence conditions (1 and 2). The participants in the deductive conditions (3 and 4) scored higher on specificity than the participants in the inductive conditions (1 and 2). This finding contrasts with hypothesis 2. In addition, participants in deductive conditions (3 and 4) had a significantly lower average case duration during the test. Therefore, the participants in the deductive conditions (3 and 4) were not only better in correctly identifying the normal images, but were also faster in their interpretation.

By contrast with hypothesis 1, no effect of instructional sequence was found on sensitivity. This analysis may have been influenced by the high test-phase sensitivity scores. As sensitivity was high in all four conditions, a ceiling effect might have occurred.

The sensitivity scores of the practice phase were lower in the inductive as well as the deductive conditions than the sensitivity scores of the test phase. In the practice phase, indeed, a significant effect in favour of the deductive conditions (3 and 4) was found.

A closer look at the results of the practice phase can provide more insights into the effects of instructional sequence on students' learning. In line with hypothesis 4, the participants in the inductive conditions (1 and 2) scored lower on sensitivity, specificity and diagnostic performance. The students in the inductive conditions were supposed to use the practice cases to develop their own solutions and were thus expected to make more mistakes during the practice phase. However, by contrast with hypothesis 4, the students in the inductive conditions (1 and 2) took less time to complete the practice cases. This suggests that they did not explore the cases in enough depth. The inductive approach might therefore not have led to productive failure during the practice phase but to *un*productive failure. Furthermore, the binary logistic regression analysis revealed that students in all four conditions needed more time for cases they incorrectly interpreted compared with cases they correctly interpreted. This finding indicates that productive failure probably occurred in all four conditions and not only in the inductive conditions (1 and 2). Invoking productive failure may therefore not be confined to inductive approaches and research on other incentives to invoke productive failure is therefore advised.

The lack of increased productive failure in the inductive conditions (1 and 2) is also reflected by the diagnostic performance scores of the test phases. No effect of sequence was found, by contrast with hypothesis 2. One of the claims for the use of inductive approaches is that they lead to deeper understanding of the problem. In this study, no evidence for this claim was found. A

deductive approach is advocated for learners who already have some experience in the task.[26] These third-year medical students can be considered novices in the task of image interpretation. However, they may already have acquired some knowledge on chest (patho)physiology during their prior medical training. This knowledge base might possibly have been solid enough for students to benefit from the deductive approach. Inductive approaches are traditionally advised for the educational experiences of learners confronted with a completely novel task.[26] Less experienced students, that is first-year medical students, might have profited from an inductive approach, and replication of this research with less experienced students is therefore advised.

A theoretical pitfall of a criterion shift used should be considered. Because of current educational practice, students are more likely to make false positive interpretations. False positive and false negative interpretations of medical images have different consequences for patient outcome. False positive interpretations may lead to unnecessary diagnostic procedures, whereas false negative interpretations may lead to potentially life-threatening delays in diagnoses.[25] However, novices generally make more false positive errors than false negative errors. This is even the case for the interpretation of images with a much lower prevalence than chest X-rays, such as the prevalence of breast abnormalities in breast cancer screening programmes.[27,28] It is unlikely that a shift in the criteria used by novices would lead to an increase in false negative interpretations. It is therefore advised to take the prevalence of diseases into account when developing training.

With the limited time that faculty members have available for medical image perception training[6], the question arises: How should students be trained to identify diverse pathologies when still developing realistic criteria?[28]Additional e-learning modules containing large image banks with the proportion of abnormalities seen in everyday clinical practice are advised.

Additionally, the use of a deductive approach is advised. In many faculties, medical image perception training is provided when students already have acquired some knowledge on anatomy and pathophysiology.[29]

Some limitations of this research are worth considering. In this research, learning outcomes were directly measured and no measures of retention of knowledge were used. Inductive sequences are also advocated to enhance retention of knowledge, yet evidence for this claim is still limited.[12] Further research to elucidate the effects of early practice is therefore recommended. Furthermore, participants were asked to make a clear distinction between normal and abnormal images, whereas everyday medical practice is not that black and white. In everyday medical practice abnormal images still predominantly consist of normal areas and normal images can contain aberrations that could be abnormal in some clinical cases. To ensure a clear cut-off between normal and abnormal in this study, only images with apparent abnormalities were used and clinical information was not provided to participants.

CONCLUSION

On immediate post-testing, a deductive approach for training third-year medical students to interpret radiographs yielded better results than an inductive approach in discerning normal from abnormal images. Furthermore, it was shown that the proportion of normal images in a training situation impacts the criteria students use to categorise normal and abnormal medical images. In many medical situations, the prevalence of disease is low and the sensitivity and specificity trade-off should be an important consideration in training design.

Contributors: KvG has been involved in every aspect of the work, has drafted the first version and was in charge of every revision round. He has written and approved the final version and acknowledges his accountability for the work. EMK is a daily supervisor to KvG and was therefore also involved in every step of the research. EMK has made great contributions to the experimental design of the study and the analysis and interpretation of the data, has critically revised the content of multiple drafts of the work, has approved the final version and acknowledges their accountability for the work. AA was a research intern for this study and was involved in the production of experimental materials, the data collection, data analysis and interpretation of the work. AA has critically revised the content of two drafts of the manuscript, has approved the final version and acknowledges his accountability for the work. SGFR is also is a daily supervisor to KvG and was involved in every step of the research, particularly in the production of experimental materials and interpretation of the work. SGFR has critically revised the content of multiple drafts of the work, has approved the final version and acknowledges their accountability for

the work. JJGvM is the principal investigator and was particularly involved in the experimental design of the work, the data analysis and data interpretation. JJGvM has critically revised the content of multiple drafts of the work, has approved the final version and acknowledges their accountability for the work.

Acknowledgements: the authors would like to thank Dr Jimmie Leppink for his statistical advice during the data analysis.

Funding: none to declare

Conflicts of interest: none to declare.

Ethical approval: ethical approval was granted by the Ethical Review Board of the Dutch Association for Education (NVMO-ERB), file number 763.

REFERENCES

1 Iglehart JK. The new era of medical imaging–progress and pitfalls. *N Engl J Med* 2006;**354** (26):2822–8.
2 Pusic MV, Andrews JS, Kessler DO, Teng DC, Pecaric MR, Ruzal-Shapiro C, Boutis K. Prevalence of abnormal cases in an image bank affects the learning of radiograph interpretation. *Med Educ* 2012;**46** (3):289–98.
3 Gegenfurtner A, Kok E, van Geel K, de Bruin A, Jarodzka H, Szulewski A, van Merriënboer JJG. The challenges of studying visual expertise in medical image diagnosis. *Med Educ* 2017;**51** (1):97–104.
4 Kok EM, van Geel K, van Merriënboer JJG, Robben SGF. What we do and do not know about teaching medical image interpretation. *Front Psychol* 2017; **8**:309.
5 Van der Gijp A, Ravesloot CJ, Jarodzka H, van der Schaaf MF, van der Schaaf IC, van Schaik JPJ, Ten Cate TJ. How visual search relates to visual diagnostic performance: a narrative systematic review of eye-tracking research in radiology. *Adv Health Sci Educ Theory Pract* 2017;**22** (3):765–87.
6 Zafar AM. Radiology: an underutilized resource for undergraduate curricula. *Med Teach* 2009; **31** (3):266.
7 Verma V, Vasudevan V, Jinnur P, Nallagatla S, Majumdar A, Arjomand F, Reminick MS. The utility of routine admission chest X-ray films on patient care. *Eur J Intern Med* 2011;**22** (3):286–8.
8 Marcolino MS, Palhares DM, Alkmim MB, Ribeiro AL. Prevalence of normal electrocardiograms in primary care patients. *Rev Assoc Med Bras (1992)* 2014;**60** (3):236–41.
9 Ng JJ, Taylor DM. Routine chest radiography in uncomplicated suspected acute coronary syndrome rarely yields significant pathology. *Emerg Med J* 2008; **25** (12):807–10.
10 Mylopoulos M, Brydges R, Woods NN, Manzone J, Schwartz DL. Preparation for future learning: a missing competency in health professions education? *Med Educ* 2016;**50** (1):115–23.
11 Van Merriënboer JJG, Sweller J. Cognitive load theory and complex learning: recent developments and future directions. *Educ Psychol Rev* 2005;**17** (2):147–77.
12 Lee HS, Anderson JR. Student learning: what has instruction got to do with it? *Annu Rev Psychol* 2013;**64**:445–69.
13 Kapur M. Productive failure. *Cogn Instr* 2008;**26** (3):379–425.
14 Soderstrom NC, Bjork RA. Learning versus performance: an integrative review. *Pers Psychol Sci* 2015;**10** (2):176–99.
15 Kapur M. Productive failure in learning math. *Cogn Sci* 2014;**38** (5):1008–22.
16 Kok EM, Jarodzka H, de Bruin ABH, BinAmir HAN, Robben SGF, van Merriënboer JJG. Systematic viewing in radiology: seeing more, missing less? *Adv Health Sci Educ Theory Pract* 2016;**21**:189–205.
17 Van Geel K, Kok EM, Dijkstra J, Robben SGF, van Merriënboer JJG. Teaching systematic viewing to final-year medical students improves systematicity but not coverage or detection of radiologic abnormalities. *J Am Coll Radiol* 2017;**14** (2):235–41.
18 Revelle W, Zinbarg RE. Coefficients alpha, beta, omega, and the glb: comments on Sijtsma. *Psychometrika* 2009;**74** (1):145–54.
19 Field AP. *Discovering Statistics using SPSS, 3rd ed.* London, UK: Sage Publications Ltd. 2009.
20 Cohen J. *Statistical Power Analysis for the Behavioral Sciences.* New York, USA: Routledge 2013.
21 Wolfe JM, Horowitz TS, Van Wert MJ, Kenner NM, Place SS, Kibbi N. Low target prevalence is a stubborn source of errors in visual search tasks. *J Exp Psychol Gen* 2007;**136** (4):623–38.
22 Treisman M, Williams TC. A theory of criterion setting with an application to sequential dependencies. *Psychol Rev* 1984;**91** (1):68–111.
23 Maddox WT. Toward a unified theory of decision criterion learning in perceptual categorization. *J Exp Anal Behav* 2002;**78** (3):567–95.
24 Krupinski EA. Current perspectives in medical image perception. *Atten Percept Psychophys* 2010;**72** (5):1205–17.
25 Berlin L. Radiologic errors and malpractice: a blurry distinction. *Am J Roentgenol* 2007;**189** (3):517–22.
26 Van Merriënboer JJG, Kirschner PA. *Ten Steps to Complex Learning: A Systematic Approach to Four-component Instructional Design.* New York, USA: Routledge 2018.
27 Miglioretti DL, Gard CC, Carney PA *et al.* When radiologists perform best: the learning curve in screening mammogram interpretation. *Radiology* 2009;**253** (3):632–40.
28 Alberdi RZ, Llanes AB, Ortega RA *et al.* Cumulative False Positive Risk (CFPR) group. Effect of radiologist experience on the risk of false-positive results in breast cancer screening programs. *Eur Radiol* 2011;**21** (10):2083–90.
29 Linaker KL. Radiology undergraduate and resident curricula: a narrative review of the literature. *J Chiropr Humanit* 2015;**22** (1):1–8.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Table S1. Characteristics of the test phase; discrimination of the cases, mean percentage correctly identified, average case duration and diagnosis per case.

Figure S1. Example of a chest x-ray used in the experiment.

Figure S2. Violin plots of the outcome measures of the test phase per condition. Violin plots represent a regular box plot with 95% confidence intervals, median and interquartile range surrounded by a rotated kernel density plot. A: Sensitivity, B: Specificity, C: Diagnostic performance and D: Average case duration.

Figure S3. Violin plots of the outcome measures of the practice phase per condition. Violin plots represent a regular box plot with 95% confidence intervals, median and interquartile range surrounded by a rotated kernel density plot. A: Sensitivity, B: Specificity, C: Diagnostic performance and D: Average case duration.

Received 27 March 2018; editorial comments to authors 7 September 2018; accepted for publication 13 September 2018