Case Comparisons: An Efficient Way of Learning Radiology

Case Comparisons:

An Efficient Way of Learning Radiology

Ellen M. Kok, MSc, Anique B. H. de Bruin, PhD, Jimmie Leppink, PhD, Jeroen J. G. van Merriënboer, PhD, Simon G. F. Robben, PhD

Rationale and Objectives: Radiologists commonly use comparison films to improve their differential diagnosis. Educational literature suggests that this technique might also be used to bolster the process of *learning* to interpret radiographs. We investigated the effectiveness of three comparison techniques in medical students, whom we invited to compare cases of the same disease (same-disease comparison), cases of different diseases (different-disease comparison), disease images with normal images (disease/normal comparison), and identical images (no comparison/control condition). Furthermore, we used eye-tracking technology to investigate which elements of the two cases were compared by the students.

Materials and Methods: We randomly assigned 84 medical students to one of four conditions and had them study different diseases on chest radiographs, while their eye movements were being measured. Thereafter, participants took two tests that measured diagnostic performance and their ability to locate diseases, respectively.

Results: Students studied most efficiently in the same-disease and different-disease comparison conditions: test 1, $F(3, 68) = 3.31$, $P = .025$, $\eta_p^2 = 0.128$; test 2, $F(3, 65) = 2.88$, $P = .043$, $\eta_p^2 = 0.117$. We found that comparisons were effected in 91% of all trials (except for the control condition). Comparisons between normal anatomy were particularly common (45.8%) in all conditions.

Conclusions: Comparing cases can be an efficient way of learning to interpret radiographs, especially when the comparison technique used is specifically tailored to the learning goal. Eye tracking provided insight into the comparison process, by showing that few comparisons were made between abnormalities, for example.

Key Words: Case comparison; eye movements; education; learning; radiology.

It is common practice for radiologists to compare films of a particular patient over time. This practice is taught to radiologist in training (1). It was found that, especially in the case of junior radiology residents, abnormalities are more easily detected when a prior image with no abnormalities (normal image) is presented alongside the case to be diagnosed (2). Hence, comparison can help to differentiate abnormalities from normal anatomy (3).

In a context of radiology education, it is of paramount importance that students *learn* to recognize common abnormalities on radiographs (4). Educational literature suggests that the use of comparison can bolster this learning process (5–8). The web-based training program COMPARE (University of Erlangen-Nuremberg, Erlangen, Germany) (5,7), for example, uses a page format in which a normal image flanks a pathologic image, and students are prompted to compare these. As much as 91% of the students and 88% of the residents who used this program valued the technique as useful or very useful (7). In addition, it was found that students learned more effectively when comparing focal diseases (ie, lesions in one location) to normal images than when comparing two pathologic images (6).

What the aforementioned studies did not probe, however, is whether such a pathologic/normal comparison technique still holds superiority in the face of a no-comparison/control condition. Besides this alternative, two other comparison options have been left uninvestigated: comparison of two images of patients with different diseases and comparison of two images of patients with the same disease. The extent to which these different comparison techniques can be effective for learning, to date, has not been investigated.

Arguably, case comparisons could be more time-consuming than a simple review of individual cases; therefore, it is important that the time spent on learning be recorded. In addition, caution should be exercised that learning materials are not presented in a suboptimal way, as this can impose an extraneous cognitive load on students' minds, that is, a cognitive load that does not contribute to learning but may hamper learning (9). Therefore, it is critical to check that the addition of a second case for comparison purposes does not inflate extraneous cognitive load. These two factors could influence

Acad Radiol 2015; 22:1226–1235

From the Department of Educational Development and Research, School of Health Professions Education, Maastricht University, PO Box 616, Maastricht, MD 6200, The Netherlands (E.M.K., A.B.H.d.B, J.L., J.J.G.v.M.); and Department of Radiology, Maastricht University Medical Center, Maastricht, The Netherlands (S.G.F.R.). Received March 5, 2015; accepted April 27, 2015. **Address correspondence to:** E.M.K. e-mail: e.kok@maastrichtuniversity.nl

http://dx.doi.org/10.1016/j.acra.2015.04.012

the extent to which case comparisons can be effective techniques for learning to interpret chest radiographs.

Another question that remains unanswered is how students avail themselves of the opportunity to compare; researchers are still in the dark about what happens during the comparison process. More specifically, we do not even know whether comparisons are actually effected when participants are presented with two or more juxtaposed images. For example, the COMPARE program instructs participants to compare the pathologic image with the normal image, but the researchers have to take it for granted that the participants actually adhere to these instructions. In such cases, eye-tracking technology (10) can provide a solution, as it measures the movements of the eye to see what a person is looking at, for how long, and in what order. As such, it can be deployed to verify and quantify the degree of comparison taking place, as well as to reveal the exact parts of the images that are being compared.

The present study has two aims. The first aim is to assess the effectiveness of three different comparison techniques in relation to a no-comparison/control condition. The second aim is to investigate which parts of the images are being compared by using eye tracking. In particular, we expect two types of comparisons to be effective for learning. First, comparing abnormalities to each other or to normal tissue could help students understand distinguishing features of abnormalities. Second, comparison of the normal tissue between two images (such as the shape of the hila in two patients) could help students learn what normal tissue looks like.

MATERIALS AND METHODS

Procedure

Participants were invited to study a series of 48 chest radiographs that were captioned with a diagnosis each and were always presented in sets of two. Participants were randomly assigned to one of four conditions in which they were asked to compare (1) cases of the same disease (same-disease condition), (2) cases of different diseases (different-disease condition), (3) disease images with normal images (disease/normal condition), and (4) identical images (no-comparison/control condition). The images were paired in accordance with the condition as follows: in the first condition, each disease case was put adjacent to a case of the same disease but pertinent to another patient; in the second condition, each disease case was paired with an image of another disease; in the third condition, each disease case was placed alongside a normal image, that is, an image showing no abnormalities; and finally in the control condition, each case was put beside an identical case, so comparison was pointless. Figure 1 showcases examples of such case pairs for each of these four conditions.

Although the participants in the first three conditions received explicit instructions to compare the two images, those in the control condition were informed about the two images being identical. All case pairs were presented in a random order and had a 30-second time slot each, but moving on to the next case pair was allowed if the participant finished earlier. The 30-second maximum was based on pilot testing.

First, the eye tracker was calibrated by repeating a 9-point calibration until accuracy was less than 1° of visual angle on both the x- and y-axis. As they had their eye movements measured, participants undertook to study the case pairs. As soon as this study phase had ended, the eye tracker was turned off. Participants subsequently indicated the extent to which they had experienced extraneous cognitive load during studying the case pairs. They then proceeded with two tests, which were identical for all participants: (1) a multiple-choice question (MCQ) test of 30 questions, which aimed to measure diagnostic performance; and (2) a region of interest (ROI) test that required participants to indicate which part of the image was abnormal by drawing an ROI around the abnormality (ROI test) to measure their ability to locate the disease. The experiment ended by thanking the participants for participation and presenting them a gift voucher.

Participants

A total of 84 third-year medical students (65 female) were participants, with mean age of 22.06 years (standard deviation, SD = 1.54). Three students were excluded from the analysis outright, as two of them reported a substantial amount of prior experience of radiology (>50 hours), and the third one had accidentally partaken in the study phase of two conditions. The 81 students that remained reported little prior experience of radiology (<2 hours) and were evenly distributed between the four conditions, with 21 participants in the same-disease condition and 20 participants in each of the other conditions. Furthermore, eye-tracking data of nine participants were excluded from the analysis as well because of insufficient data quality (ie, during calibration, the threshold of 1° of visual angle could not be reached). Eventually, the analysis of eye-tracking data included 20 participants in the same-disease condition, 17 in the different-disease condition, 16 in the disease/normal condition, and 19 in the control condition.

Cases

Although the term "case" is usually taken to denote the ensemble of one or more radiographs, patient history, and clinical questions for the purpose of this experiment, we use this term to refer to individual posterioranterior (PA) chest radiographs void of any additional information. For each of eight different diseases, a board-certified radiologist collected nine cases with a typical radiographic manifestation. The final diagnosis was established based on clinical information, clinical course, and other images (eg, computed tomography or chest radiographs made at other moments). Four of these diseases were focal in kind (atelectasis, solitary lung tumor, pneumonia, and pleural effusion), that is, the abnormality was centered in one location with the rest of the lung being normal (11), whereas the other four were diffuse diseases, in which the whole lung was abnormal (cystic fibrosis, lung fibrosis, metastases, and

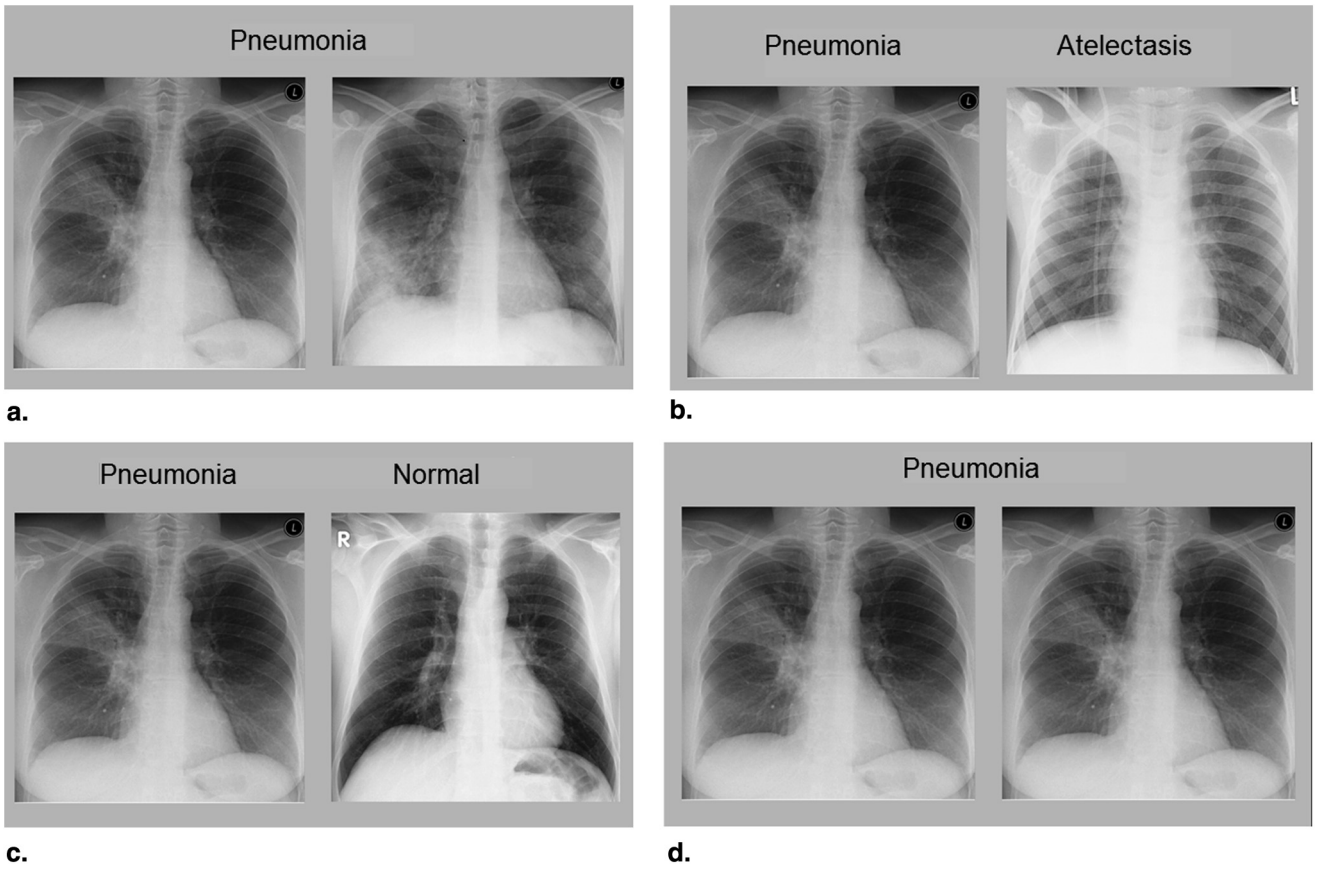

Figure 1. Screenshots of the study phase: **(a)** same-disease condition; **(b)** different-disease condition; **(c)** disease/normal condition; and **(d)** control condition. Names of diseases have been translated from Dutch.

miliary t*uberculosis*). Six of each set of nine cases were destined for use in the study phase and three cases were intended for the test. The 24 test cases that resulted (three times eight disease cases) were subsequently complemented by six normal cases, so the number of test cases totaled 30; the study phase contained 48 cases (six times eight disease cases), supplemented by an additional set of 14 normal images in the disease/normal condition. See Figure 2 for an overview of the cases in each phase of the experiment. All images were stripped of any identifying information, and resized to be 800 pixels in height (width differed between images). Cases were presented on a computer monitor and captioned with the correct diagnosis only.

To keep all other things constant such that eye-tracking data could be adequately compared, participants in the control condition were not presented with one but two identical cases. Yet, they were informed about the cases being identical before the start of the experiment.

Measures

Performance Test. Performance was assessed by means of two consecutive tests, which participants were allowed to take at their own pace. The two tests aimed to capture two different aspects of chest radiograph interpretation: the ability to diagnose the disease, and the ability to locate the disease. In both tests, single cases, not case pairs were presented. The first test, an MCQ test of 30 questions, aimed to measure diagnostic performance. With each question, participants were shown a single case void of any information and asked which disease was visible. In answering, participants could choose one from a list of the aforementioned eight diseases, or "no disease." A separate MCQ test score was computed for both the disease cases and the normal cases, each score representing the percentage of correct answers.

The second test, which aimed to measure participants' ability to locate the disease, provided participants with the same cases as those of the MCQ test, but did give a diagnosis. Normal cases were excluded. They were asked to draw an ROI around that part of the image they deemed abnormal by using the mouse. The region drawn by two thorax radiologists was then compared to the participant's drawing and the percentage of overlap was calculated. The aggregate score was the average percentage of overlap.

Cognitive Load. Ineffective learning can be the result of extraneous cognitive load that is generated by a suboptimal presentation of the learning task (9). To ascertain that the comparison techniques used would not impose a high extraneous cognitive load on learners because of bad design, we measured extraneous cognitive load by means of an extraneous load scale that forms part of an existing and validated cognitive load inventory (12). This 10-point scale consisted of three questions.

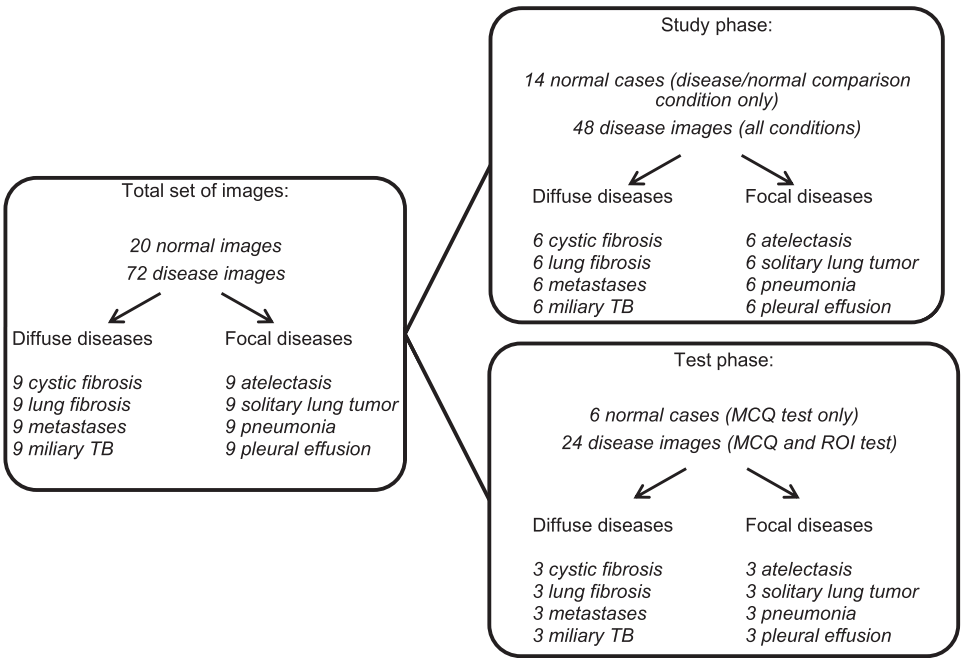

Figure 2. Overview of the cases used and their assignment to the phases of the experiment. TB, tuberculosis.

With the maximum score being 10, throughout this article the average of ratings is reported.

Apparatus

Eye movements were gauged by means of an SensoMotoric Instruments RED 250 eye tracker (www.smivision.com). The study phase of the experiment was prepared and executed using SensoMotoric Instruments Experiment Center software (www.smivision.com). The MCQ test was created in E-Prime (http://www.pstnet.com/eprime.cfm), and the ROI test was presented in CAMPUS (http://www.medizinische-fakultaet-hd.uni-heidelberg.de/CAMPUS-Software.109992.0.html).

Analyses

One-way analyses of variance (ANOVAs) were performed to test intergroup differences between the means of the four conditions for all dependent variables. For post hoc analyses, a Bonferroni correction was applied, so the adapted alpha was 0.05/6 = 0.008. As to the ANOVAs, effect size 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. Because the effect size for the overall ANOVA gives less information than the effect sizes for individual comparisons (13), we used Cohen's d to qualify the differences found in the post hoc tests, with 0.2 being considered a small effect, 0.5 a moderate effect, and 0.8 a large effect (14).

Eye-tracking data were collected at 250 Hz. The minimum fixation duration was set at 50 milliseconds. A saccade is a rapid eye movement during which no information is taken as given in Ref. (10). Each saccade that started in one of the images and landed in the other image (transition) was regarded as a "comparison saccade."

Because eye tracking yields enormous amounts of data, it was not feasible to perform a detailed analysis of all data. Therefore, we took a subset of the eye-tracking data from the three comparison conditions and analyzed it in more detail to investigate which elements of the cases were compared by the students. To this end, we randomly selected 60 trials of focal cases and 60 trials of diffuse cases, which were each stratified for condition such that the analysis included 40 trials from each comparison condition. All comparison saccades were classified into three groups: (1) comparison involving a focal abnormality (either starting or ending in an abnormality, or both), (2) comparison of the same organ (starting and ending in the same organ, but in different images; these saccades were mainly horizontal saccades), and (3) comparison of different organs (ending in a different organ than the one it started in).

RESULTS

Test Scores

All test results are displayed in Table 1. A moderate correlation between the MCQ test and the ROI test scores was found, $r = 0.308$ and $P = .006$.

MCQ Test: Disease Cases. The average score of one of the disease questions in the MCQ test correlated negatively with the average total score for disease images, thereby violating the assumption that additional questions contribute positively to the reliability of the average total score. After removing this question, the Cronbach's alpha improved

TABLE 1. Average Scores and Standard Deviations for the Four Conditions on the MCQ Test (Disease and Normal Questions Separately), ROI Test, Extraneous Cognitive Load Scale, and Time Spent Learning

	MCQ Test (Disease Cases)		MCQ Test (Normal Cases)		ROI Test		Extraneous Cognitive Load		Time Spent Learning (min)	
Condition	M	SD	M	SD	M	SD	M	SD	M	SD
Disease/normal comparison	11.3 (49.1%)	3.4	2.8 (45.8%)	1.5	30.2%	11.4	0.5	0.7	9.0	3.0
Same-disease comparison	12.5 (54.3%)	2.6	1.3 (22.1%)	1.2	34.8%	6.8	1.0	1.3	7.8	2.7
Different-disease comparison	13.6 (59.1%)	2.7	2.1 (35.0%)	1.6	34.5%	12.3	0.7	0.8	8.5	2.5
No comparison	13.5 (58.7%)	2.7	2.0 (33.3%)	1.8	33.6%	10.5	1.0	1.2	11.5	4.3

M, mean; MCQ, multiple-choice questions; ROI, region of interest; SD, standard deviation.

The MCQ scores are expressed as number of cases correctly identified, with its related percentage in parentheses. The ROI test score is the percentage of overlap. The extraneous cognitive load is the average score (maximum score is 10).

from 0.50 to 0.54, whereas none of the remaining questions were negatively correlated to the average total score. The MCQ test scores did not reveal any significant effect of condition, $F(3, 77) = 1.60$, $P = .20$, $\eta_p^2 = 0.059$.

MCQ Test: Normal Cases. The MCQs about "normal" images (showing no abnormalities) were analyzed separately. The six normal questions together had a Cronbach's alpha of 0.57. A significant effect of condition on number of images correctly identified as normal was found, $F(3, 77) = 3.01$, $P = .035$, $\eta_p^2 = 0.105$ (see Table 1). Post hoc analyses indicate that participants in the disease/normal condition were more successful in distinguishing normal from abnormal cases than participants in the same-disease condition ($P = .004$, Cohen's $d = 1.08$). However, we found no significant difference between the disease/normal condition and both the different-disease condition ($P = .18$, Cohen's $d = 0.44$) and the control condition ($P = .12$, Cohen's $d = 0.47$). The different-disease condition did not differ significantly from the same-disease condition ($P = .11$, Cohen's $d = 0.57$), nor from the control condition ($P = .84$, Cohen's $d = 0.06$). Finally, the latter two groups did not differ significantly between them ($P = .16$, Cohen's $d = 0.45$).

ROI Test. Cases showing metastases were removed from the analyses of the ROI test, because when drawing ROIs around each metastasis, many of the participants halted as soon as they noticed that many metastases were visible and communicated this verbally instead. With those images removed, the Cronbach's alpha of the ROI test was 0.82. None of the separate average scores correlated negatively with the average total score. No significant effect of condition was found on the ROI test scores, $F(3, 73) = 0.26$, $P = .86$, $\eta_p^2 = 0.010$.

Extraneous Cognitive Load

The Cronbach's alpha for the extraneous cognitive load scale was 0.54. This value is somewhat lower than values found in previous studies (12), which indicates that it might be attributable to the restricted range in extraneous cognitive load scores (ie, most participants rated the extraneous cognitive load as low on all three questions, see Table 1). From these data, we can infer that the chosen form of presenting the learning material constituted no further impediment. No significant differences were found between conditions, $F(3,77) = 1.12$, $P = .35$, $\eta_p^2 = 0.042$.

Time Spent Studying

We gauged differences between the four groups in the time they needed to study all pair of images, hereinafter referred to as "dwell time." The differences were significant, $F(3, 68) = 4.66$, $P < .005$, $\eta_p^2 = 0.181$ (see Table 1). Post hoc analyses revealed that total dwell time was significantly higher in the control condition compared to both the different-disease condition ($P = .008$, Cohen's $d = 0.86$) and same-disease condition ($P = .001$, Cohen's $d = 1.05$). The total dwell time in the disease/normal condition did not differ significantly from any of the other conditions (different-disease comparison, $P = .63$, Cohen's $d = 0.21$; same-disease comparison, $P = .25$, Cohen's $d = 0.45$; control condition, $P = .031$, Cohen's $d = 0.66$). Neither could we establish any significant differences between the different-disease and same-disease conditions ($P = .51$, Cohen's $d = 0.28$).

Efficiency

Because of the great variability in dwell time, we calculated an efficiency measure for the MCQ test (disease cases) and the ROI test, which factors in the time spent studying (efficiency = [z-test score − z-study time]/$\sqrt{2}$) (15). In this sense, "efficiency" denotes a state in which test result and the time taken up in the study phase are inversely related: efficiency increases as test results become higher and the time spent studying lessens and vice versa.

For both the MCQ test (disease items) and the ROI test, significant differences between conditions in efficiency were found [MCQ, $F(3, 68) = 3.31$, $P = .025$, $\eta_p^2 = 0.128$; ROI, $F(3, 65) = 2.88$, $P = .043$, $\eta_p^2 = 0.117$], with the different-disease and same-disease conditions ranking highest (see Fig 3 and Table 2). After a post hoc test with adjusted alpha was conducted, however, none of the differences reached significance (see Table 3). Only the same-disease condition revealed a marginally significant advantage over the control condition on the ROI test.

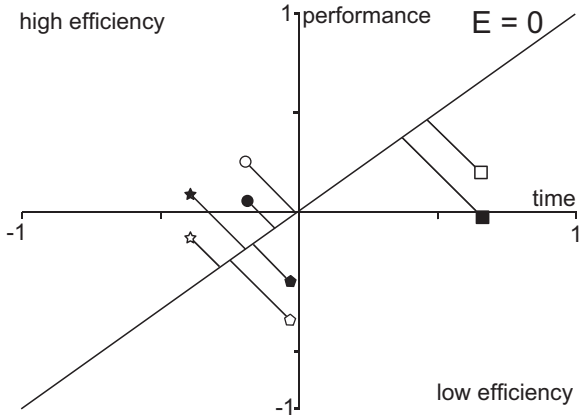

Figure 3. Efficiency for the control condition (□ and ■), disease/normal condition (⬠ and ⬟), different-disease condition (○ and ●), and same-disease condition (☆and ★). Filled figures represent the MCQ test z-score, and open figures represent the ROI test z-score. Because time spent learning refers to the study phase of the experiment, the MCQ test and ROI test scores of each condition have identical values on the x-axis. The diagonal line labeled $E = 0$ indicates an efficiency of zero. Lines extending to the upper left corner indicated increased efficiency; lines extending to the lower right corner indicated decreased efficiency. See Ref. (16) for more information about the efficiency plot. MCQ, multiple-choice questions; ROI, region of interest.

Eye Tracking

Eye-tracking data were collected in the study phase only. In the three comparison conditions, participants made at least two comparison saccades in 91% of all trials. The average number of comparison saccades per trial was 7.30 (SD = 5.5) in the disease/normal condition, 6.90 (SD = 4.3) in the same-disease condition, and 5.09 (SD = 4.6) in the different-disease condition. There was no correlation between the average number of comparison saccades and performance (all three comparison conditions pooled: ROI test, $r = -0.190$, $P = .19$; MCQ test, $r = -0.051$, $P = .72$). As for the control condition, at least two comparison saccades were made in 55% of all trials. Although participants were not instructed to compare, an average of 2.49 (SD = 3.0) comparison saccades were effected anyway. Figure 4 showcases some typical comparison scan paths.

The comparison saccades of 120 randomly selected trials from the comparison conditions were subjected to further scrutiny to understand what elements of the two cases were compared by the students (see Table 4). In these trials, a total of 639 comparison saccades were effected, which translates to a slightly more than six comparisons per trial on average. All comparison saccades were classified as being (1) a comparison involving an abnormality, (2) a comparison of the same organ, or (3) a comparison of different organs. Comparisons that either start or end in an abnormality in one of the images, or both, are labeled as "comparison involving an abnormality" (eg, starting in the lung in the left image and ending in a tumor in the right image). Although comparisons involving an abnormality were quite common in the different-disease and same-disease conditions, they resulted not so in the disease/normal condition. Comparisons that both began and ended in a focal abnormality were mainly found in the same-disease condition ($n = 11$); only one such comparison was found in the different-disease condition, and of course these were not possible in the disease/normal condition. Comparison saccades involving an abnormality were more likely to start in an abnormality ($n = 42$) than to end in an abnormality ($n = 21$). By extension, five of those ending in an abnormality were immediately followed by a saccade that started in that abnormality. Comparisons that started in one of the images and ended in the same organ of the other image were labeled as "comparisons of the same organ" (eg, mediastinum in the left image with mediastinum in the right image). These could be found in almost half of the trials. Such comparisons were mostly effected between the heart, lungs, mediastinum, hila, or abdomen of the two images. Comparisons that ended in a different organ than the one it started in were labeled "comparisons of different organs" (eg, between the heart in the left image and the lung in right image). In general, they were slightly less common than the comparisons of the same organ.

TABLE 2. Average Efficiency for the Four Conditions on the MCQ Test and ROI Test

	Efficiency MCQ Test		Efficiency ROI Test	
Condition	M	SD	M	SD
Disease/normal comparison	−0.36	0.71	−0.18	0.85
Same-disease comparison	0.18	0.76	0.33	0.61
Different-disease comparison	0.32	0.89	0.18	1.04
No comparison	−0.32	0.81	−0.51	1.19

M, mean; SD, standard deviation.
Efficiency = (z-test score − z-study time)/√2.

DISCUSSION

The present study has sought to assess the effectiveness of three different comparison techniques in relation to a no-comparison control condition: comparison to a normal image (disease/normal condition), comparison of cases of the same disease (same-disease condition), and comparison of cases of different diseases (different-disease condition). Average scores of the students for both the diagnostic performance test (MCQ test, disease cases) and the ROI test that measured the ability to locate the disease did not appear to differ over conditions. Peculiarly, we did find that participants in the disease/normal condition correctly identified a larger number of normal cases. The presupposition that the use of comparison would impose on students a higher extraneous cognitive load, luckily, could not be confirmed.

Although we did not find significant differences in test scores between conditions, the conditions did vary markedly

TABLE 3. *P*-Values for Post Hoc Tests for Efficiency

	Efficiency MCQ Test			Efficiency ROI Test		
Post-hoc comparisons	Mean Difference	*P* Value	Cohen's *d*	Mean Difference	*P* Value	Cohen's *d*
Disease/normal comparison						
Same-disease comparison	−0.54	.047	0.75	−0.52	.119	0.74
Different-disease comparison	−0.68	.017	0.86	−0.36	.289	0.39
No comparison	−0.04	.895	0.05	0.33	.326	0.32
Same-disease comparison						
Different-disease comparison	−0.14	.607	0.17	0.16	.619	0.19
No comparison	0.51	.052	0.66	0.84	.009	0.93
Different-disease comparison						
No comparison	0.64	.019	0.78	0.69	.036	0.63

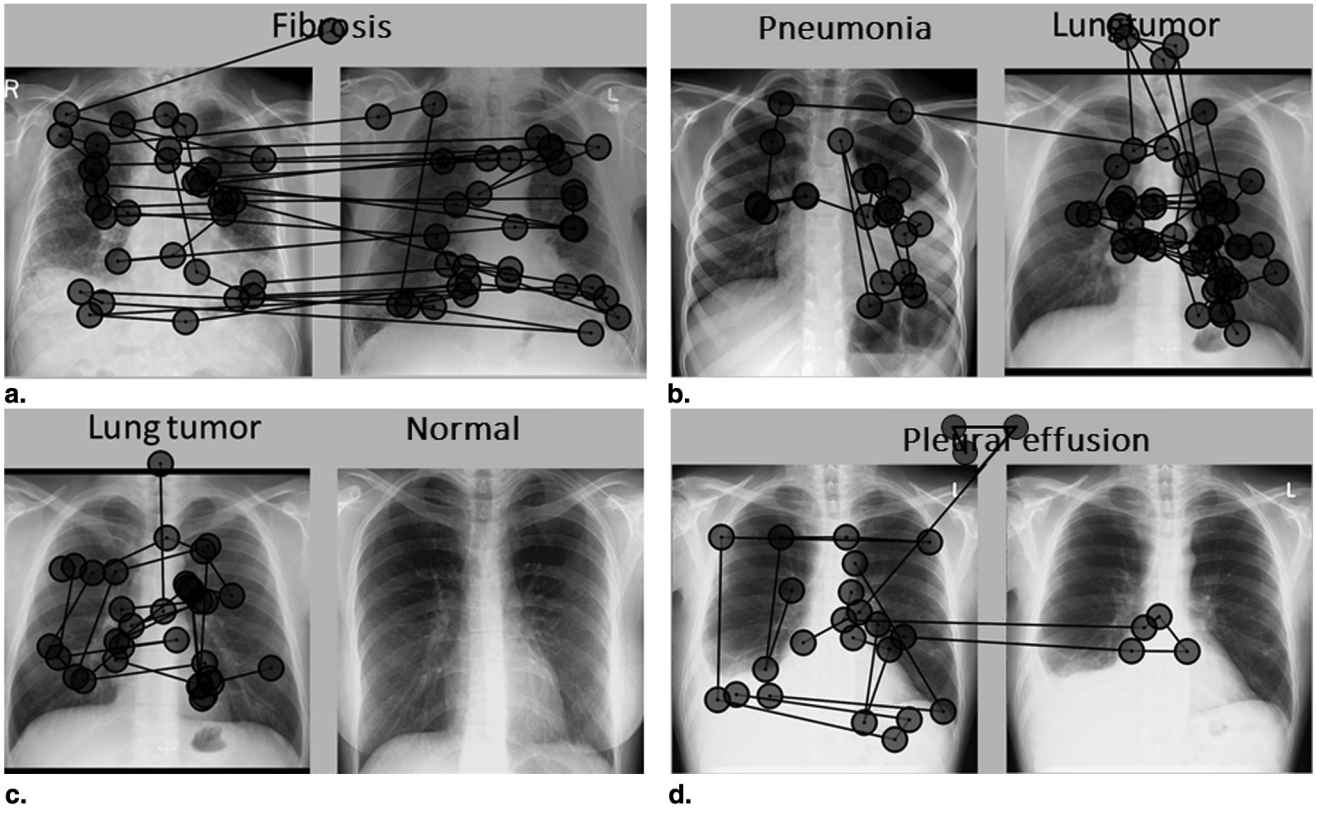

Figure 4. Scan paths of four different trials: **(a)** a participant in the same-disease condition (two cases of fibrosis) who makes many comparisons; **(b)** a participant in the different-disease condition (comparison of pneumonia with a tumor) who works in a sequential manner; **(c)** a participant in the disease/normal condition (comparison of lung tumor with a normal image) who ignores the normal image; and **(d)** a participant in the control condition (two identical images of a patient with pleural effusion) who makes two comparison saccades to the identical image on the right, but mainly focuses on the left image. Names of diseases have been translated from Dutch.

with respect to the time participants needed to study the images. Especially, the participants in the control condition required almost 30% more time to study the cases compared with the other three conditions.

By including time in the calculation of efficiency, we found that the highest levels of efficiency were attained when same-disease and different-disease comparison techniques were used: when on both tests participants performed similarly or even better with respect to the other two conditions and required less time. It is important to note that participants in these two conditions had not been exposed to more pathology: all participants reviewed the same number of pathology cases; hence, the increased efficiency must have been attributable to the opportunity to compare cases.

In terms of efficiency, the group effecting same-disease comparisons performed best on the ROI test, whereas the opposite was true for the group comparing different diseases, which performed best on the MCQ test. These findings

TABLE 4. Classification of 639 Comparison Saccades From 120 Randomly Selected Trials, Showing Which Elements of the Cases Were Compared by the Students

	Study Condition						
	Disease/Normal Condition		Different-Disease Condition		Same-Disease Condition		Total
Type of Comparison	Focal (*n* = 20)	Diffuse (*n* = 20)	Focal (*n* = 20)	Diffuse (*n* = 20)	Focal (*n* = 20)	Diffuse (*n* = 20)	(*n* = 120)
(1) Involves an abnormality	12 (9.8 %)		24 (27.0 %)		39 (31.2 %)		75 (11.7 %)
(2) Comparison of the same organ	49 (39.8 %)	65 (63.1%)	36 (40.4 %)	46 (52.3 %)	40 (32.0 %)	57 (51.3 %)	293 (45.8 %)
(3) Comparison of different organs	62 (50.4 %)	38 (36.9%)	29 (32.4 %)	42 (47.7 %)	46 (36.8 %)	54 (48.3 %)	271 (42.4 %)
Total number of comparison saccades	123 (100 %)	103 (100 %)	89 (100 %)	88 (100 %)	125 (100 %)	111 (100 %)	639 (100 %)

A trial refers to the eye movements of one participant on one case pair. Forty trials from each condition (20 focal case pairs, 20 diffuse case pairs) were randomly selected. All comparison saccades in these trials (639 in total) were classified as (1) a comparison involving an abnormality, (2) a comparison of the same organ, or (3) a comparison of different organs. Comparisons in the control condition have not been analyzed. Numbers and percentages add up to 100% vertically, representing the total number of saccades affected in the 20 trials within a condition and type of image. For example, of all 123 saccades affected in the 20 focal trials from the disease/normal condition, 12 (9.8%) were comparisons involving an abnormality, 49 (39.8%) were comparisons of the same organ, and 62 (50.4%) were comparisons of different organs.

resound the contention of Hammer et al. (17) that, in general, comparison of things that are different (in this case radiographs of different diseases) can help a student to identify and learn their discriminating features. This is reflected in the MCQ test that measured diagnostic performance, because being able to distinguish between different diseases is central to good performance on this test.

Comparison of things that belong to the same category (like different patients with the same disease), on the other hand, can help discover the different manifestations of a disease (17). Use of same-disease comparisons could be the best technique to teach a student about the ranges of pneumonia and their differences. It seems plausible that students in the same-disease comparison condition gained a better insight into the variation within a disease, which in turn helped them to detect the borders and size of an abnormality, and, consequently, to localize the abnormality. Understanding the range within which a disease can manifest itself is important in deciding which part of the image is normal, and which is not.

Participants in the disease/normal condition performed best with respect to the identification of normal cases in the MCQ test. As they were the only group to have been exposed to normal images during the study phase, the opportunity to compare these images to disease images might have helped them to learn the distinction between the two. At the same time, however, this comparison technique proved less efficient at learning the distinction between different diseases and the variation within the disease, as scores were relatively low for both the MCQ (with regard to disease cases) and ROI tests.

In summary, application of different comparison techniques led to equally different emphases on different elements of learning to interpret radiographs. Although disease/normal comparisons seemed most effective at learning to discriminate between normal and abnormal images, use of different-disease comparisons seemed the most appropriate technique for learning the distinction between different diseases. By the same token, same-disease comparisons seemed most effective at understanding the different manifestations of a disease and the range of the disease. Thus, it seems important that learners deploy such comparison techniques strategically, as the learning activity's objective requires. So simply put, the distinction between pneumonia and cystic fibrosis should not be taught by having the learner sequentially compare both diseases with normal images, but by having them compare the diseases with each other. The sequential comparison of those cases to a normal image, however, could help students to learn to differentiate between normal and abnormal images.

To expand on the aforementioned, we postulate that students can gain most from comparison techniques if presented in a specific order: (1) the disease/normal comparison technique, so that they learn to differentiate between normal and abnormal images; (2) the different-disease comparison technique, by which they learn to distinguish one disease from another; and (3) the same-disease comparison technique, to teach them the different manifestations of a particular disease. However, further research is required in which the effectiveness of teaching in this specific order is studied.

The second aim was to investigate which parts of the images are being compared, by using eye tracking. The eye-tracking data threw more light on this. What stood out was that participants really did avail themselves of the opportunity to compare. Moreover, they often compared between normal anatomy, such as the shape of the hila or the size of the mediastinum between the two cases.

Comparison of the abnormality with normal tissue in the juxtaposed case was less common. This was surprising because educational literature suggests that comparison of an abnormality with normal tissue or with another abnormality could help students understand the distinguishing features of the abnormality (17). For example, comparison of the heart border

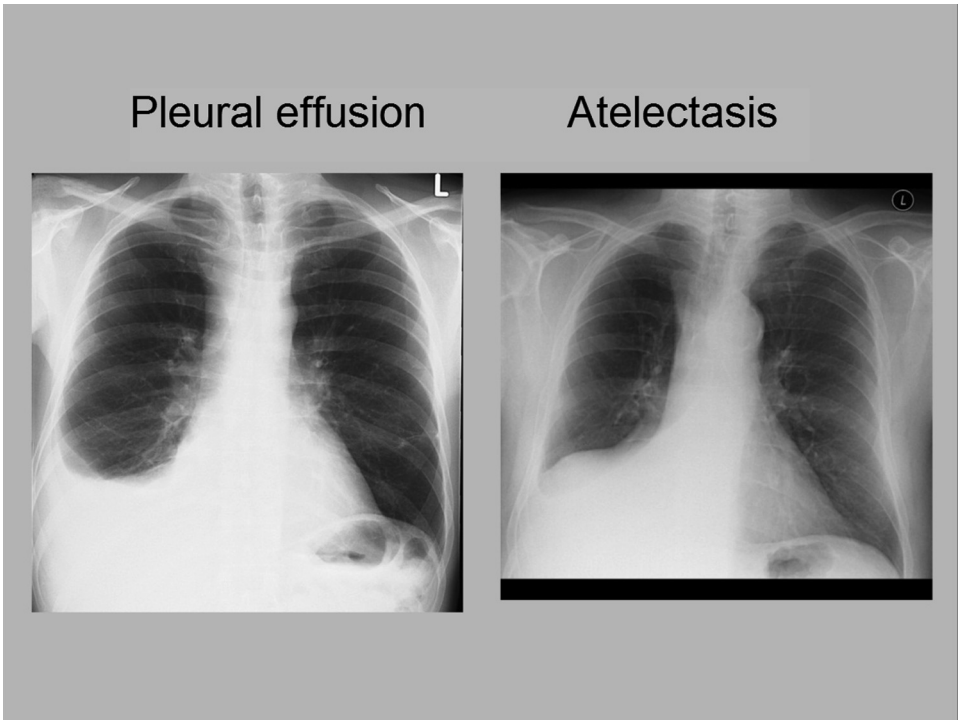

Figure 5. Screenshot from the different-disease condition, showing a pleural effusion in the left image and an atelectasis in the right image. Names of diseases have been translated from Dutch.

of a patient with pneumonia with the heart border of a healthy patient could help the student understand the silhouette sign in pneumonia. Therefore, if we want students to compare abnormalities with normal tissue, they need explicit instructions to do so (18). Students inexperienced in radiology might also need us to direct their attention to the abnormalities, as they could have difficulty detecting these (19).

One of the limitations of this experiment is that the eye-tracking data were observational. Had participants received explicit instructions as how to go about comparing, then the effect on performance of the comparison method used could have been investigated outright. Instructing students to compare normal with abnormal tissue, or to focus on similarities or differences between cases, for example, could have rendered comparison techniques more effective for learning. However, the advantage of our observational method is that we could see how and what people compare when given the choice. Prospective studies could focus on whether provision of different sets of comparison instructions could influence the effectiveness of specific comparison techniques.

It should not be difficult to implement case comparisons in different teaching settings, as they can be easily introduced into lectures and case reviews, for example. They meet the requirement of Gunderman et al. (20) that teaching should not be just about delivering concrete facts but also include higher-order concepts. Case comparisons could move a learner beyond a mere understanding of what an abnormality looks like in a single case toward understanding how higher-order concepts are expressed in different patients and different diseases. For example, comparison might help inexperienced medical students understand that pleural effusion is characterized by a concave surface, which distinguishes it from, for instance, a basal atelectasis (see Fig 5).

A second limitation of the current experiment is that it was confined to a population of inexperienced medical students and to the educational realm of chest radiography. We do believe, however, that there is a scope for more seasoned medical students and residents to benefit from such comparison techniques too, provided the degree of case complexity is raised. In addition to this, we also expect that the principle can be generalized to other image modalities and anatomic regions as well, although further research is required to investigate whether the principles found generalize to other modalities, in particular to multiplanar images such as computed tomography and magnetic resonance imaging. Effective use of case comparisons requires an extensive teaching file, so a teacher or a student can quickly look up matching cases of a disease or relevant cases of a different disease.

CONCLUSIONS

Our study has demonstrated that, compared to the "traditional" disease/normal case comparisons, alternative comparison techniques are equally or even more effective. Eye-tracking data confirm that students indeed do compare cases when given the opportunity.

ACKNOWLEDGMENTS

The authors would like to thank Angelique van der Heuvel for editing the manuscript.

REFERENCES

1. Carmody DP, Kundel HL, Toto LC. Comparison scans while reading chest images. Taught, but not practiced. Invest Radiol 1984; 19: 462–466.
2. Berbaum KS, Franken EA, Jr, Smith TJ. The effect of comparison films upon resident interpretation of pediatric chest radiographs. Invest Radiol 1985; 20:124–128.
3. Carmody DP, Nodine CF, Kundel HL. Finding lung nodules with and without comparative visual scanning. Percept Psychophys 1981; 29: 594–598.
4. Kondo KL, Swerdlow M. Medical student radiology curriculum: what skills do residency program directors believe are essential for medical students to attain? Acad Radiol 2013; 20:263–271.
5. Grunewald M, Heckemann RA, Gebhard H, et al. COMPARE radiology: creating an interactive Web-based training program for radiology with multimedia authoring software. Acad Radiol 2003; 10:543–553.
6. Kok EM, de Bruin ABH, Robben SGF, et al. Learning radiological appearances of diseases, does comparison help? Learn Instr 2013; 23:90–97.
7. Wagner M, Heckemann RA, Nomayr A, et al. COMPARE/radiology, an interactive Web-based radiology teaching program: evaluation of user response. Acad Radiol 2005; 12:752–760.
8. Hatala RM, Brooks LR, Norman GR. Practice makes perfect: the critical role of mixed practice in the acquisition of ECG interpretation skills. Adv Health Sci Educ Theory Pract 2003; 8:17–26.
9. van Merrienboer JJG, Sweller J. Cognitive load theory in health professional education: design principles and strategies. Med Educ 2010; 44: 85–93.
10. Holmqvist K, Nyström M, Andersson R, et al. Eye tracking: a comprehensive guide to methods and measures. Oxford: Oxford University Press, 2011.
11. Kok EM, de Bruin ABH, Robben SGF, et al. Looking in the same manner but seeing it differently: Bottom-up and expertise effects in radiology. Appl Cognitive Psych 2012; 26:854–862.
12. Leppink J, Paas F, van Gog T, et al. Effects of pairs of problems and examples on task performance and different types of cognitive load. Learn Instruct 2014; 30:32–42.
13. Field AP. Discovering statistics using SPSS: London: Sage Publications, 2009.
14. Cohen J. Statistical power analysis for the behavioral sciences. Hillsdale, NJ: Lawrence Earlbaum Associates, 1988.
15. Van Gog T, Paas F. Instructional efficiency: revisiting the original construct in educational research. Educ Psychol 2008; 43:16–26.
16. Paas FG, Van Merriënboer JJ. The efficiency of instructional conditions: an approach to combine mental effort and performance measures. Human Factors 1993; 35:737–743.
17. Hammer R, Diesendruck G, Weinshall D, et al. The development of category learning strategies: what makes the difference? Cognition 2009; 112:105–119.
18. Alfieri L, Nokes-Malach TJ, Schunn CD. Learning through case comparisons: a meta-analytic review. Educ Psychol 2013; 48:87–113.
19. Reingold EM, Sheridan H. Eye movements and visual expertise in chess and medicine. In: Leversedge SP, Gilchrist ID, Everling S, eds. Oxford handbook on eye movements. Oxford: Oxford University Press, 2011; 528–550.
20. Gunderman R, Williamson K, Fraley R, et al. Expertise: implications for radiological education. Acad Radiol 2001; 8:1252.