Title
How effective is perceptual training? Evaluating two perceptual training methods on a difficult visual categorisation task

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 43(43)

ISSN
1069-7977

Authors
Marris, Jessica
Perfors, Andrew
Mitchell, David
et al.

Publication Date
2021

Peer reviewed

How effective is perceptual training? Evaluating two perceptual training methods on a difficult visual categorisation task

Jessica Marris (j.marris@student.unimelb.edu.au)
School of Psychological Sciences, University of Melbourne, Melbourne, Australia

Andrew Perfors (andrew.perfors@unimelb.edu.au)
School of Psychological Sciences, University of Melbourne, Melbourne, Australia

David Mitchell (dpm.mitchell@gmail.com)
Radiology, Sligo University Hospital, Sligo Ireland
Department of Radiology, The Royal Melbourne Hospital, Melbourne, Australia

Wayland Wang (wayland.wang@mh.org.au)
Department of Radiology, The Royal Melbourne Hospital, Melbourne, Australia

Mark W. McCusker (mark.mccusker@mh.org.au)
Department of Radiology, The Royal Melbourne Hospital, Melbourne, Australia
Department of Radiology, University of Melbourne, Melbourne, Australia

Timothy John Haynes Lovell (tim.lovell@mh.org.au)
Department of Radiology, The Royal Melbourne Hospital, Melbourne, Australia

Robert N. Gibson (robert.gibson@mh.org.au)
Department of Radiology, The Royal Melbourne Hospital, Melbourne, Australia
Department of Radiology, University of Melbourne, Melbourne, Australia

Frank Gaillard (frank.gaillard@mh.org.au)
Department of Radiology, University of Melbourne, Melbourne, Australia
Department of Radiology, The Royal Melbourne Hospital, Melbourne, Australia

Piers Douglas Howe (pdhowe@unimelb.edu.au)
School of Psychological Sciences, University of Melbourne, Melbourne, Australia

Abstract

Perceptual training leads to improvements in a wide range of simple visual tasks. However, it is still unclear how effective it can be for more difficult visual tasks in real-world domains such as radiology. Is it possible to train people to the level of experts? If so, what method is best, and how much training is necessary? Over four training sessions, we trained medically naïve participants to identify the degree of fatty liver tissue present in ultrasound images. We found that both COMPARISON and SINGLE-CASE perceptual training techniques resulted in significant post-training improvement, but that the SINGLE-CASE training was more effective. Whilst people showed rapid learning with less than one hour of training, they did not improve to the level of experts, and additional training sessions did not provide significant benefits beyond the initial session. This suggests that perceptual training could usefully augment, but not replace, the traditional rule-based training that medical students currently receive.

Keywords: perceptual training; comparison; radiology; expertise

Introduction

Perceptual learning is defined as the improvement in task performance that occurs with sensory experience (Sagi, 2011). It has been shown to occur across a range of low-level visual tasks including motion direction detection (Ball & Sekuler, 1987), orientation discrimination (Fiorentini & Berardi, 1980), and texture discrimination (Karni & Sagi, 1991), using simple stimuli such as random dot motion, line segments, and Gabor patches. Perceptual training is based on pattern recognition, which enables learners to become sensitive to important features and relationships even when these relationships are difficult to verbalise (Kellman & Garrigan, 2009).

It has been shown that perceptual training can result in a substantial improvement in performance for a variety of real-world visual tasks in medical domains such as radiology (Chen, HolcDorf, McCusker, Gaillard, & Howe, 2017; Frank et al., 2020; Johnston et al., 2020; Sha, Toh, Remington, & Jiang, 2020; Sowden, Davies, & Roling, 2000), dermatology (Xu, Rourke, Robinson, & Tanaka, 2016), histopathology (Krasne, Hillman, Kellman, & Drake, 2013), and cytology (Evered, Walker, Watt, & Perham, 2014). Because some aspects of perceptual expertise are difficult to express verbally, perceptual training has been proposed as a potential supplement to the traditional approach to training medical professionals like radiologists, who are currently trained to diagnose medical images in a primarily rule-based fashion (Johnston

et al., 2020). However, it remains unclear how much benefit people can receive from perceptual training: can it lead to expert-level performance in highly complex domains?

While many studies have demonstrated that perceptual training results in rapid learning, few studies have compared the performance of participants trained by perceptual training to that of experts. One exception is a study by Chen et al. (2017), in which medically naïve participants were trained to identify hip fractures in X-ray images. After only two training sessions (1280 training images in total), the mean accuracy of novices was approximately 90%, which was slightly lower than experts (radiology residents and board-certified radiologists, who achieved approximately 94% accuracy). However, the top five novices performed at a level comparable to the board-certified radiologists after less than one hour of training. These findings support the idea that perceptual training can result in a level of expertise that is practically useful, and thus could usefully augment the traditional rule-based training paradigm.

An alternative possibility is that the task in Chen et al. (2017) may have been relatively easy for novices to learn and perceptual training would have been less effective for a more difficult task. Consistent with this, in a more difficult task that involved identifying whether appendicitis was present in a single axial image from a computed tomography scan, Johnston et al. (2020) found that experts substantially outperformed trained novices. Consequently, it remains unclear whether perceptual training can lead to similar levels of performance as experts on more difficult radiological tasks.

A second unresolved issue is which perceptual training methodology is the most effective. The standard approach, which has been used by the majority of studies, involves presenting a single stimulus on each trial. The participant answers a specific question about the image (e.g., "Is there a hip fracture in this image?") and receives immediate feedback. This method was used by both Chen et al. (2017) and Johnston et al. (2020), along with studies involving the categorisation of cancerous lesions (Xu et al., 2016) and the recognition of histopathology patterns (Krasne et al., 2013).

An alternative training method, which we refer to as comparison training, presents multiple stimuli simultaneously on each trial. These stimuli generally depict different categories (e.g., a normal and abnormal medical image) and require an alternative forced choice response (e.g., "Which image is abnormal?"), which is followed by feedback. While only a few perceptual training studies have used comparison training (Evered et al., 2014; Sha et al., 2020), similar techniques are widely used in category learning experiments (e.g., Kang & Pashler, 2012; Meagher, Goldstone, Nosofsky, & Carvalho, 2017).

The proposed advantage of comparison training is that it enhances discriminative contrast by highlighting commonalities and differences (Kang & Pashler, 2012). Comparing stimuli that represent differing categories can improve discrimination ability (Hammer, Bar-Hillel, Hertz, Weinshall, & Hochstein, 2008), particularly when discriminating between highly similar categories (Carvalho & Goldstone, 2014). Despite the potential benefits of comparison training, to our knowledge, there are no perceptual learning studies with complex real-world stimuli that have compared comparison training to standard (single-case) perceptual training.

A third unresolved issue is the amount of training required for perceptual learning with complex visual stimuli. Most studies have used multiple sessions (generally between two and four), which is consistent with the positive effects of distributed practice (Donovan & Radosevich, 1999). However, it is unclear whether multiple sessions are, in fact, necessary. Several previous studies used relatively short training sessions because these studies were limited by the number of stimuli they had access to. Given that repeating images within a training session does not appear to reduce the effectiveness of the training (Chen et al., 2017; Sha et al., 2020), these studies might have achieved the same result using just a single, but longer, training session. Alternatively, it may be that while fewer training sessions is adequate for relatively simple images, more sessions may be required when the domain or task is more challenging.

The current study was designed to address these three issues. First, we specifically chose a task that trainee radiologists find difficult – identifying the degree of fatty liver disease present in ultrasound images – as this is the type of task where experts are likely to outperform non-experts, and compared the performance of our trained participants to that of experts. Second, we evaluated performance under two different types of perceptual training techniques, COMPARISON and SINGLE-CASE training. Third, we investigated the learning benefits of multiple training sessions.

We hypothesised that both types of perceptual training regimes would lead to an improvement in post-training performance, but that COMPARISON training would be more effective because the task requires discriminating between very similar images. Due to the large literature on distributed practice (Donovan & Radosevich, 1999), we further hypothesised that multiple training sessions would result in more learning beyond that of a single training session but, regardless of the extent of training, we would not be able to train medically naïve participants to the level of experts on such a difficult task.

Method

Participants

We recruited 186 medically naïve adults from Prolific Academic to complete a four session experiment online. A prescreening questionnaire was used to ensure that all participants had normal-or-corrected-to-normal visual acuity, normal colour vision, and no prior experience in radiology. Participants were compensated for each session, receiving £11.05 for completing all sessions. Twenty participants did not complete all sessions and five were excluded due to repeating a session, resulting in 161 people in the dataset (89

female; M_{age} = 37.35 years; SD_{age} = 12.60 years; all resided in the UK, US, Canada, Australia, or New Zealand).

We additionally recruited a group of experts; three consultant radiologists, one radiology fellow, and one radiology registrar from the Royal Melbourne Hospital. These experts did not participate in the experiments but instead graded the stimuli and provided a measure of expert performance.

Materials

The stimuli were abdominal ultrasounds of 505 unique livers obtained from a tertiary care centre. Each liver was represented by a collage with four ultrasound views of it (Figure 1). We used a collage of four images to depict each case instead of a single image, because in practice, radiologists make decisions about these types of cases based on several images.

As there is no objective measure to determine the degree of fatty liver tissue, we used ratings from our experts to establish a gold standard consensus grade. The experts independently reviewed each collage and graded the degree of fatty liver disease (*hepatic steatosis*) present, ranging from 1 (Normal) to 7 (Severe). The average of all of the five experts was used as the gold standard for training and assessing naïve participants. Because we wished select stimuli that were rated consistently by the experts, we excluded cases where any expert's rating was more than one off the consensus grade. This resulted in a stimulus pool of 386 unique collage images for use in the experiment.

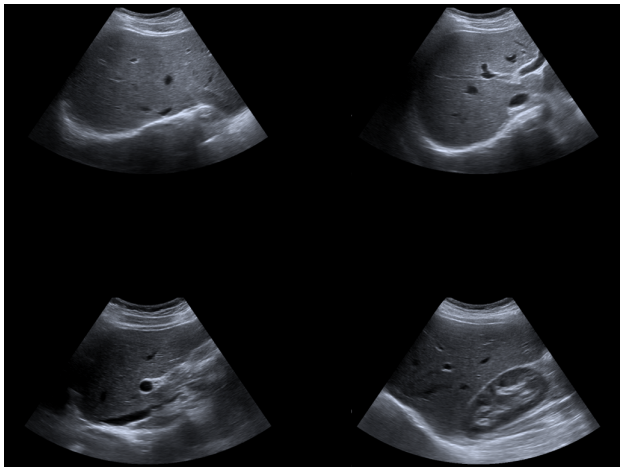

Figure 1: Example of a collage of two transverse and two longitudinal liver ultrasound images from a single case. In this example, the degree of fatty liver disease is 1 (Normal).

The collages were split into a training set and two test sets (286 collages in the training set and 50 collages in each test set) such that the distribution of grades was balanced between each set. There was no overlap between the training and test sets and the same sets were used across all participants.

The experiment was created using jsPsych (de Leeuw, 2015) and participants completed it online using a laptop or desktop computer; they were prevented from using tablets and smartphones. Each collage was resized to a width of 750 pixels and a height of 562.5 pixels. A minimum browser window of 1024x700 pixels was required.

Design and Procedure

We used a pre-test/post-test design, with two possible perceptual training conditions (between-participants; COMPARISON and SINGLE-CASE, described below). For technical reasons, the single-case training condition was run after the comparison training condition. In both training conditions, the experiment occurred over four sessions, with demographic questions and a pre-training test at the beginning of the first session, perceptual training in sessions 1-4, and a post-training test at the end of the fourth session. Participants were provided with a 48 hour window to complete each session, which was followed by a 24 hour break before the next session began. Thus, depending on when during the 48 hour window the participant completed the session, the participant would experience a break of between 24 and 120 hours between sessions. All of the sessions were self-paced, with no limits on the time taken to view and respond on each trial.

At the start of each session, participants were shown a brief explanation of the grading scale with some example images and an annotated image that identified the main differences between normal and severe fatty liver disease. All participants then completed the pre-test, where they were asked to grade the degree of fatty liver tissue present (using the 7-point grading scale described previously) for 50 unique collage images (Figure 1 provides an example of a collage image for a single-case). The collage images were presented sequentially (a single case per trial) in a randomised order. Participants used the keyboard to grade each collage according to the 7-point grading scale. No feedback was provided during the pre-test.

There were 100 training trials per session for both of the training conditions (400 trials total). Images were randomly sampled from the training set and could be repeated across sessions or even within a session. To motivate participants throughout the training, points were awarded based on performance. Participants were able to monitor and compare their points across sessions.

After completing the final training phase in session four, all participants completed a post-test. The format of the post-test was the same as the pre-test, except the 50 collage images were novel.

Single-case Training On each trial in the SINGLE-CASE perceptual training (n = 90 participants), participants were presented with a collage consisting of four images of the same case, and were asked to rate the degree of fatty liver disease on the 7-point scale described above. This was a similar format to the pre-test and post-test. However, after they provided their assessment, the correct grade was immediately presented underneath the image for review. In addition, a feedback message was presented in coloured text, with the content of the message determined by how close the response was to the correct answer ("Spot on! Correct" in green for

Figure 2: Example of a COMPARISON training trial. In each trial, participants were presented with two panels of images. Each panel contained a single liver case, with each case depicting a different degree of fatty liver tissue. From the two livers, participants were asked to make a discrimination choice. In this example, the left panel depicts a grade of 5 (Moderate), the right panel depicts a grade of 3 (Mild), and participants were asked which panel was Mild.

correct, "Almost" in blue for one grade from the correct answer, "Not quite" in orange for two grades from the correct answer, or "Incorrect" in red for responses more than two grades from the correct answer).

Comparison Training In the COMPARISON training (n = 71 participants), people were asked to compare two different livers. They were thus presented with two panels of images (each consisting of half of a collage image from a single case) simultaneously on each trial. We chose to present half of each collage in order to keep the total number of individual images of the livers (i.e., four) consistent with the SINGLE-CASE training condition. The two panels were positioned side-by-side with a small gap between them (Figure 2). Each panel depicted a different grade of fatty liver tissue.

The training task in the COMPARISON condition differed from the SINGLE-CASE condition, as we asked people to discriminate between two different grades of fatty liver tissue. At the start of each trial, a prompt informed participants about the discrimination that they would need to make (e.g., "Which image is 3: Mild?"). The prompt also contained information about the grade of the other image (e.g., "The other image is 5: Moderate"). Participants indicated which image they believed matched the prompt and then received immediate feedback about whether they were correct or not.

The difficulty of the comparisons was adapted throughout the training, and began with easier comparisons (i.e., cases that were six grades apart). The difficulty of the comparison is determined by the distance between the grades of the livers being compared.

A modified 2-up 1-down adaptive staircase procedure was used, whereby correct responses on two previous trials stepped a participant up to the next difficulty level (e.g., from six grades apart to five apart, and so on). Following an incorrect discrimination, participants returned to a lower difficulty comparison. The adaptive nature of the training meant that participants progressed up and down, depending on their performance.

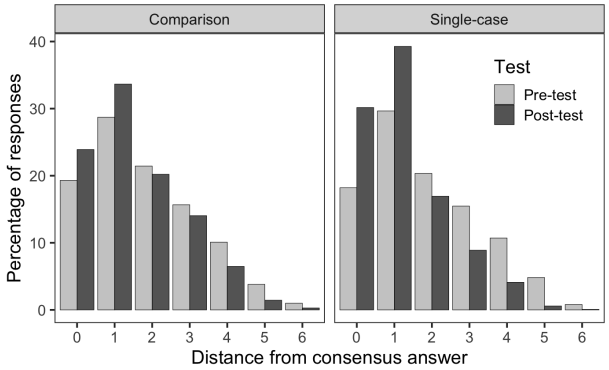

Figure 3: Percentage of responses for each possible distance from the consensus answer in the pre-test and post-test for each training condition. A distance of 0 is equivalent to a correct response (consistent with the consensus answer). In both conditions, people gave more correct or near-correct answers in the post-test, indicating that they learned something.

Results

Due to a technical error, data from 28 trials across seven participants was missing, but the remaining data from these participants was retained. The dataset and analyses can be found at `https://github.com/jmarris/pt-livers`. People in the SINGLE-CASE training took an average of 9 minutes to complete each session, while those in the COMPARISON training took 12.

Figure 3 shows a comparison of pre-test to post-test performance across both training conditions. It is evident that in both conditions, people were more accurate in the post-test, making more responses that were closer to the correct answer (distances 0 or 1) and fewer that were further (distances 5 or 6). This suggests that there was some learning between pre-test and post-test.

In order to better quantify overall improvements in performance as well as compare across conditions, we calculated each participant's mean error on each test, which represents their mean distance from the consensus answer. This is shown in Figure 4. For the sake of comparison, we also show the performance of our group of experts on the same images that the medically naïve participants were tested on in the pre-test and post-test. For these experts, we needed to use a slightly different reference point than that used to assess performance of the naïve participants. To avoid 'double-dipping' the data, we assessed the performance of each expert in turn, by calculating their performance relative to the other four experts. From this, we calculated an overall measure of expert performance, which is shown in the green bars. It is evident that although both types of perceptual training were successful for training medically naïve participants, as indicated by the post-test improvement, the training alone was not sufficient for them to reach similar levels of performance as the group of experts.

Figure 4: Performance by training condition on the pre-test and post-test images for medically naïve participants. The experts did not undergo training, but provide a measure of expert performance on the same images. The y axis shows the mean error (distance from the consensus answer); thus, lower is better. Each dot is the mean error for one individual, and error bars represent the standard error. Both training conditions improved from pre-test to post-test, but improvement was greater for SINGLE-CASE training.

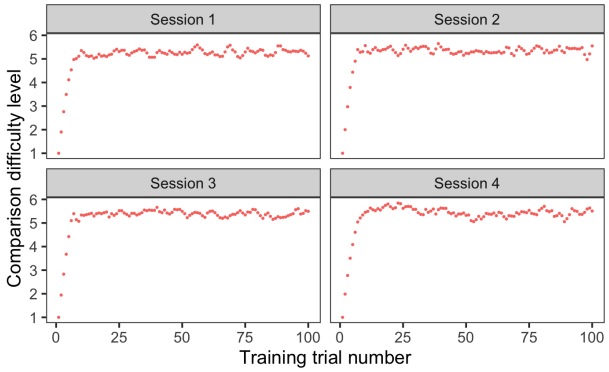

Figure 5: Performance improvement across time during COMPARISON training. Each panel corresponds to a session, and the x axis indicates the trial number within session. The y axis reflects the difficulty level of the comparison, with 6 being the most difficult and 1 being the easiest. Participants improve rapidly within each session as the algorithm adapts to them, but there is no discernible improvement across sessions; people converge to a difficulty level of five in all, which corresponds to a difference of two grades between comparison stimuli.

In order to quantify whether performance varied by training condition or test, we performed a mixed 2x2 ANOVA on all of the non-expert data. Training condition was a between-subjects factor (COMPARISON or SINGLE-CASE), test phase (PRE or POST) was a within-subjects factor, and mean error was the dependent variable. As expected, there was a significant main effect of the test phase, $F(1,159) = 196.79, p < .001, \eta^2_G = .340$, with lower mean error in the post-test compared to the pre-test indicating successful learning.

There was also a significant main effect of training condition, $F(1,159) = 9.74, p = .002, \eta^2_G = .035$ and a significant interaction $F(1,159) = 24.74, p < .001, \eta^2_G = .061$. This suggests that the effect of training was greater in the SINGLE-CASE condition. Post-hoc t-tests with Bonferroni corrections supported this interpretation. There was no significant difference in performance between the training conditions on the pre-test, $t(148.28) = -0.72, p = .944$, but on the post-test participants who received SINGLE-CASE training had significantly lower error than people who received COMPARISON training, $t(139.38) = 6.56, p < .001$.

How did learning proceed over the course of training? Is it possible that medically naïve participants could have reached the level of experts had we provided them with only a few additional sessions? Figures 5 and 6 address these questions by presenting the training data broken down by session.

For the COMPARISON training in Figure 5, mean error is not a good measure of performance since the training approach is adaptive; we therefore examine the difficulty level of the comparison (difficulty is 7 minus the distance between grades, so that 1 is the easiest and 6 is the most difficult). It is evident that each session is marked by rapid increase in the difficulty of the comparisons as the algorithm adapts to each participant; however, performance across sessions does not appear to improve, with people converging on a difficulty level of around five (i.e., discriminating between livers that are two grades apart) in all sessions. To address whether the additional three training sessions provided a benefit above what learning was achieved in the first session, we compared the mean difficulty level on the final training trial in the first training session and last training session. Evaluating performance on the final trial enables us to compare performance after the learning from the first session has occurred with performance after the learning from all sessions has occurred. A paired-samples t-test revealed no significant difference, $t(70) = -1.48, p = .145$, suggesting that all of the meaningful improvement in learning between the pre-test and the post-test likely occurred during the first session.

For the SINGLE-CASE training in Figure 6, there appears to be a slight decrease in error over the course of the first training session, but not in subsequent sessions. We quantified this with a linear model for each session, using error as the outcome variable and trial number as the predictor. It was statistically significant for the first session, $F(1,98) = 22.77, p < .001$, but not for any of the other three. Consistent with this, a paired-samples t-test revealed no significant difference in the mean error on the final training trial in the first session and final session, $t(89) = 1.42, p = .160$. Taken together, these results suggest that additional training would not help participants reach the level of expert performance, at least not within a small number of additional sessions.

Discussion

Our work makes several contributions to the growing literature investigating perceptual training in real-world, visually complex domains. First, we replicated the finding that perceptual training improves the classification of medical images, as seen by the significant reduction in mean error on the post-training test; less than one hour of training enabled our participants to generalise their learning to novel images.

This is particularly interesting as we were able to train medically naïve participants to do a task that trainee radiologists find difficult – identifying the degree of fatty liver tissue in ultrasound images. Whilst our perceptual training was not as successful as Chen et al. (2017), it is not surprising given the higher difficulty of the perceptual discrimination required in our task. However, consistent with Johnston et al. (2020), perceptual training alone was not sufficient to train people to the level of experts.

How much does the type of training matter? As two different methods have been used in the literature, it was unclear whether one might be superior. Given the potential advantages of comparison training (e.g., Hammer et al., 2008; Kang & Pashler, 2012), we expected that discriminating between different grades of fatty liver tissue would enhance learning, because the stimuli are highly similar and differences only become obvious when the stimuli are viewed side-by-side. However, we found that SINGLE-CASE perceptual training resulted in better generalisation.

One explanation for this finding is that transfer is better when the task people are tested on aligns closely with the training task. Another possibility is that the SINGLE-CASE training allowed people to learn about the full underlying distribution of different grades of fatty liver tissue in our experiment, since this distribution was the same in both the training and test sets. Conversely, since comparison training was adaptive, the distribution of the grades of fatty disease observed by participants would not necessarily have matched the distribution of grades seen in the test images. Whilst participants in the COMPARISON training would have seen more livers overall (unique and repeated cases), only half of the full collage of four images for each case was presented on each trial. This method was chosen to ensure a consistent number of images were presented in each training condition. Additionally, it avoided overcrowding the display further, which could have impacted on learning (e.g., due to higher cognitive load).

We realise that our training method differs from the procedure commonly used in the perceptual training literature (i.e., presenting a single image or a pair of images), as we presented multiple images on each trial in both conditions (i.e., the collages). This change was because radiologists generally make these decisions from multiple ultrasound images. Whilst it is unclear whether our methodological change impacts on the effectiveness of perceptual training, we show that perceptual learning occurs when perceptual decisions are based on multiple simultaneously presented images.

Finally, we found that learning was rapid and that multiple sessions did not provide substantial benefits to learning, suggesting that we can achieve the benefits of perceptual training within just one training session. In the COMPARISON training, people rapidly progressed to more difficult comparisons within a session but this pattern did not substantially change over sessions; performance after four training sessions was not significantly better than after one session. The SINGLE-CASE training also showed a similar lack of benefit for multiple sessions. For both perceptual training conditions, rapid initial learning occurred within one session, but learning hit a ceiling after this point. While it is theoretically possible that a sufficient number of additional sessions might allow people to improve to expert levels, the trend suggests that it would take a very large number. It is also possible that for complex domains such as this, where successful identification relies on theoretical knowledge, purely perceptual training alone cannot produce expert-level classification.

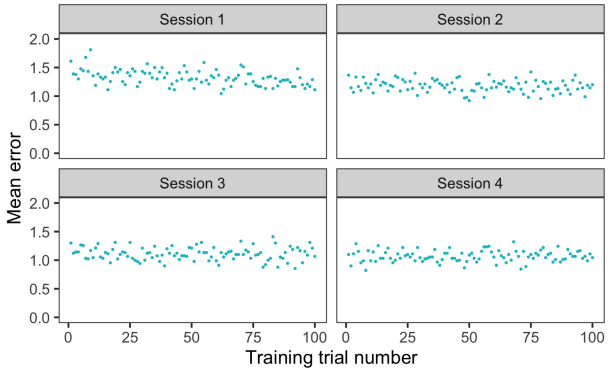

Figure 6: Performance improvement across time during SINGLE-CASE training. Each panel corresponds to a session, and the x axis indicates the trial number within session. The y axis reflects the mean error. Participants show gradual improvement over the first session, but no noticeable improvement in subsequent sessions.

It could be suggested that our comparison of medically naïve participants and experts is not fair because the experts were evaluated on their own ratings. However, to limit 'double-dipping' we evaluated each expert relative to the other experts, and found that they were all quite consistent in their ratings. We believe that the comparison to experts is still useful, as it provides a benchmark for how effective our training paradigm is.

An avenue for future work is investigating if there are ways to improve the effectiveness of perceptual training in difficult tasks. Is it possible to boost learning further? Recent work by Frank et al. (2020) suggested that perceptual learning might be improved if it is augmented with detailed feedback during training. In their work, supplementing partial feedback (e.g., "Response correct: Lesion is present" and "Identified location of the lesion is incorrect") with information about the location (with annotations) was necessary for perceptual learning of both calcification and distortion lesions in mammograms. Additionally, people that received detailed feedback showed long-term retention when tested six months later, but people who only received partial feedback did not. This finding could explain the success of the training in Chen et al. (2017), who used annotated feedback that directed attention to the location of the bone fracture.

Whilst the possibility of augmenting perceptual training with detailed feedback provides a promising avenue for future work, there are some practical considerations. It can be

time consuming, costly, and difficult for experts to annotate or label images. Additionally, for some diagnoses, it may be difficult to isolate and communicate the particular regions of interest to people that are medically naïve. One potential solution, which we plan to investigate in our future work, is to test the benefit of providing limited numbers of annotated images with verbal "perceptual rules" via augmented perceptual training.

Our findings have implications for future applications of perceptual training. We demonstrate that perceptual training can be practically useful and efficient: even in a complex real-world domain, people can significantly and rapidly improve their performance with less than one hour of training. Although people did not reach expert performance, we show that perceptual training is a tool that can be used to achieve rapid initial learning, which can be subsequently refined further. A combined approach would offer multiple benefits. Perceptual training would provide trainees with the experience of a large number of exemplars, including those that may not be often seen in practice (Johnston et al., 2020). This would provide a robust foundation for subsequent rule-based learning and clinical reasoning.

Acknowledgments

This research was funded by a Royal Australian and New Zealand College of Radiology research grant (grant number 20187/RANZCR/011). JM was supported by an Australian Government Research Training Program Scholarship.

References

Ball, K., & Sekuler, R. (1987). Direction-specific improvement in motion discrimination. *Vision Research*, *27*(6), 953–965.

Carvalho, P. F., & Goldstone, R. L. (2014). Putting category learning in order: Category structure and temporal arrangement affect the benefit of interleaved over blocked study. *Memory & Cognition*, *42*(3), 481–495.

Chen, W., HolcDorf, D., McCusker, M. W., Gaillard, F., & Howe, P. D. L. (2017). Perceptual training to improve hip fracture identification in conventional radiographs. *PLoS ONE*, *12*(12), 1–11.

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12.

Donovan, J. J., & Radosevich, D. J. (1999). A meta-analytic review of the distribution of practice effect: Now you see it, now you don't. *Journal of Applied Psychology*, *84*(5), 795–805.

Evered, A., Walker, D., Watt, A. A., & Perham, N. (2014). Untutored discrimination training on paired cell images influences visual learning in cytopathology. *Cancer Cytopathology*, *122*(3), 200–210.

Fiorentini, A., & Berardi, N. (1980). Perceptual learning specific for orientation and spatial frequency. *Nature*, *287*(5777), 43–44.

Frank, S. M., Qi, A., Ravasio, D., Sasaki, Y., Rosen, E. L., & Watanabe, T. (2020). Supervised learning occurs in visual perceptual learning of complex natural images. *Current Biology*, *30*(15), 2995–3000.

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.

Johnston, I. A., Ji, M., Cochrane, A., Demko, Z., Robbins, J. B., Stephenson, J. W., & Green, C. S. (2020). Perceptual learning of appendicitis diagnosis in radiological images. *Journal of Vision*, *20*(8), 1–17.

Kang, S. H., & Pashler, H. (2012). Learning painting styles: Spacing is advantageous when it promotes discriminative contrast. *Applied Cognitive Psychology*, *26*(1), 97–103.

Karni, A., & Sagi, D. (1991). Where practice makes perfect in texture discrimination: Evidence for primary visual cortex plasticity. *Proceedings of the National Academy of Sciences*, *88*(11), 4966–4970.

Kellman, P. J., & Garrigan, P. (2009). Perceptual learning and human expertise. *Physics of Life Reviews*, *6*(2), 53–84.

Krasne, S., Hillman, J. D., Kellman, P. J., & Drake, T. A. (2013). Applying perceptual and adaptive learning techniques for teaching introductory histopathology. *Journal of Pathology Informatics*, *4*(1), 297–304.

Meagher, B. J., Goldstone, R. L., Nosofsky, R. M., & Carvalho, P. F. (2017). Organized simultaneous displays facilitate learning of complex natural science categories. *Psychonomic Bulletin & Review*, *24*(6), 1987–1994.

Sagi, D. (2011). Review: Perceptual learning in vision research. *Vision Research*, *51*, 1552–1566.

Sha, L. Z., Toh, Y. N., Remington, R. W., & Jiang, Y. V. (2020). Perceptual learning in the identification of lung cancer in chest radiographs. *Cognitive Research: Principles and Implications*, *5*(1), 1–13.

Sowden, P. T., Davies, I. R. L., & Roling, P. (2000). Perceptual learning of the detection of features in X-Ray images: A functional role for improvements in adults' visual sensitivity? *Journal of Experimental Psychology: Human Perception and Performance*, *26*(1), 379–390.

Xu, B., Rourke, L., Robinson, J. K., & Tanaka, J. W. (2016). Training melanoma detection in photographs using the perceptual expertise training approach. *Applied Cognitive Psychology*, *30*(5), 750–756.