Testing the Effectiveness of Augmenting Perceptual Training With Annotations and Steps in a Difficult Visual Discrimination Task

Jessica Marris[1], Andrew Perfors[1], Robert N. Gibson[2,3], Frank Gaillard[3,2], Piers D. L. Howe[1]

[1]School of Psychological Sciences, University of Melbourne, Australia
[2]Department of Radiology, The Royal Melbourne Hospital, Melbourne, Australia
[3]Department of Radiology, University of Melbourne, Australia

Abstract

Perceptual training has been shown to be an effective and rapid way of training people to make simple diagnoses using medical images. However, it appears to be less effective at training people to make more complex diagnoses that require non-binary judgements. In the present study, we investigated whether perceptual training could be augmented to make it more effective and what factors limited its effectiveness. In Experiment 1, we created artificial stimuli that were designed to simulate liver ultrasound images to assess perceptual learning for a complex task that involved judgements on a 7-point scale. Whilst performance improved somewhat with training, we found that incorporating annotations into the training provided no benefits. Additionally, contrary to our expectations, training that was structured in a stepped fashion was detrimental to learning. In Experiment 2, we found that perceptual learning in a simple task with shaded disks was most impacted by the extent to which the brightness levels of each disk were discriminable but that attending to multiple locations did not result in a significant cost to performance. Our findings show that augmenting perceptual training does not increase learning and that learning is less when the relevant features are harder to discriminate.

Keywords: perceptual learning; categorization; learning; medical image interpretation

Introduction

Perceptual training has been found to rapidly improve performance across a number of tasks in the medical domain that involve detecting the presence/absence of targets, such as tumours in chest radiographs (Sha, Toh, Remington, & Jiang, 2020), hip fractures in X-ray images (Chen, HolcDorf, McCusker, Gaillard, & Howe, 2017), lesions in mammograms (Frank et al., 2020), and appendicitis in computed tomography images (Johnston et al., 2020). However, it appears that perceptual training is less effective for tasks that require more than a binary present/absent judgment (Marris et al., 2023). It is possible that such tasks may not benefit from perceptual training to the same extent as binary tasks because performing at a high level in these more complex tasks may require background knowledge that is better acquired via explicit, didactic instruction.

In this paper, we investigate whether the effectiveness of perceptual training for these more complex tasks can be increased by incorporating an explicit, didactic element. For instance, adding annotations that identify the target locations to the training images. The provision of annotated feedback has been found to enhance learning and generalisation in tasks that involved identifying if appendicitis was present in CT images (Johnston et al., 2020) and identifying the presence of lesions and their location in mammograms (Frank et al., 2020).

A related approach in the categorisation literature is *feature highlighting*, which involves giving learners verbal feature descriptions that are designed to direct attention to the relevant features (Meagher, McDaniel, & Nosofsky, 2022; Miyatsu, Gouravajhala, Nosofsky, & McDaniel, 2019).

An approach that has been used widely in the education domain for aiding the learning of complex tasks is to break the task into a sequence of simpler steps (Van Merriënboer, Kirschner, & Kester, 2003). This approach offers the benefit of requiring little additional effort or cost to implement while allowing the trainer to make explicit the sequence of steps the learner should follow to achieve the desired outcome.

Our aim was to investigate the efficacy of augmenting perceptual training with annotations and a stepped-learning procedure for a difficult task that requires non-binary judgements. We chose to investigate this with a task that both experts and trainee radiologists find difficult: identifying the degree (on a 7-point scale) of hepatic steatosis (fatty infiltration of the liver) that is present in ultrasound images. However, as these real-world stimuli are noisy, as an initial step, we decided to use artificial stimuli that were designed to simulate liver ultrasound images. This approach eliminated noise and allowed us to carefully construct the stimuli in a way that ensured we could rigorously assess the perceptual training paradigms that we studied. Demonstrating which perceptual training techniques are (or are not) successful with artificial stimuli will provide a useful starting point. Subsequent work will then investigate to what extent our findings can be extended to actual liver ultrasound images.

Experiment 1

In a task that simulates identifying the degree of hepatic steatosis in ultrasound images, we hypothesised that augmenting perceptual training with annotated feedback and structuring the training in a stepped manner would improve learning, as measured by a reduction in mean error between a pre-test and a post-test. Additionally, we hypothesised that both annotated feedback and a stepped training approach on their own will be more effective than a standard perceptual training procedure that does not include either annotations or steps.

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

Method

Participants We recruited 206 participants from Amazon Mechanical Turk (MTurk). Participants were compensated $4.8 and those that scored in the top 20% were awarded a bonus of $1. All reported normal-or-corrected-to-normal visual acuity, normal colour vision, and no prior experience in radiology. Data for six participants were excluded (three for failing attention checks and three due to technical issues), consistent with our pre-registered exclusion criteria (`https://aspredicted.org/N3F_C3Q`). The final dataset included 200 participants (85 females, 112 males, and 2 non-binary; M_{age} = 42.3 years; SD_{age} = 11.9 years). All participants resided in the USA, Canada, or the UK.

Materials The artificial stimuli were created to align to seven grades of hepatic steatosis, ranging from 1 (Normal) to 7 (Severe). The distribution of grades was matched to a sample of real cases that were collected from a tertiary care centre.

In consultation with domain experts, three perceptual features relevant to identifying the degree of hepatic steatosis were identified: (1) the brightness of the background liver tissue in the upper part of the image (Background), (2) the brightness of the white lines around the blood vessels (Lines), and (3) the difference in brightness between the lower and upper liver tissue (Gradient). Using plain language, brief verbal descriptions were created to describe these features and how they could be used diagnostically to identify the degree of hepatic steatosis (Table 1).

Because in practice radiologists view multiple ultrasound images of a liver when making diagnostic decisions, each training and test image comprised a collage of four liver ultrasound images (see Figure 1 for an example). We created 190 unique collages, which were split into a training set (90 collages) and two test sets (50 collages each), such that the distribution of grades was balanced between each set. Each collage was constructed to perfectly align with the description given in Table 1 (i.e. there was no noise and each feature had perfect diagnosticity). In each collage, nine to twelve unique small shapes were dispersed across the four liver images to simulate blood vessels, and the brightness of the lines around these shapes was altered to be consistent with the feature descriptions. Similarly, the brightness of the top of the liver images varied on two levels and the brightness of the bottom of the liver images image could either be the same as the top of the images, slightly darker than the top of the images or much darker than the top of the images, thereby producing a brightness gradient. To create the annotated feedback, feature descriptions were combined with circles and arrows that identified parts of the liver images that were relevant for assessing each feature.

Design and Procedure Participants were randomly assigned to one of four training conditions in a 2 (Annotations vs No Annotations) x 2 (Steps vs No Steps) design. The experiment was developed with jsPysch (de Leeuw, 2015). The

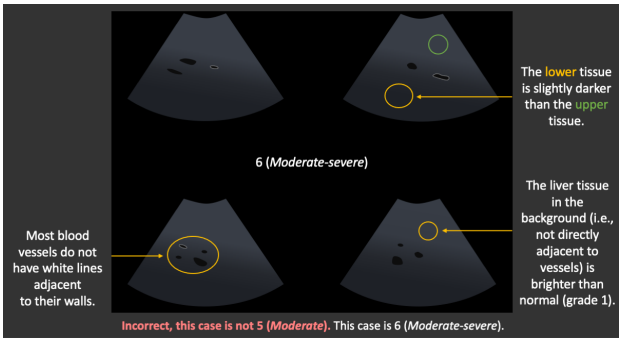

Figure 1: An example of a liver collage with annotated feedback. The annotations and feature descriptions were only provided to participants in the ANNOTATIONS conditions during the training phase, following an incorrect response. In the NO ANNOTATIONS conditions, participants only received feedback regarding the correctness of their response and the correct grade of the collage. In this example, the degree of hepatic steatosis is 6 (Moderate-severe).

experiment was self-paced and completed online. At the start, participants were informed that they were to grade each collage according to a 7-point grading scale and were provided with four examples of individual livers that depicted grades 1, 3, 5, and 7 (see Figure 2). After completing an understanding check, participants underwent a pre-test phase where they graded 50 collages without feedback (all participants were tested on the same set of collages). Following this, participants underwent perceptual training (90 trials in total; 30 trials per block) and then completed a post-test (50 trials). The post-test was the same format as the pre-test, except participants were tested on a set of 50 new collages. Participants saw the same 190 collages in all conditions.

In the NO STEPS AND NO ANNOTATIONS condition, the training was equivalent to a standard form of perceptual training and involved grading each collage on the 7-point scale and then immediately being informed of the correct grade but with no other feedback. In the training phase of the two ANNOTATIONS conditions, feedback on incorrect trials was supplemented with annotations that described how the three features could be used to determine the grade of hepatic steatosis in that instance.

In the two STEPS conditions, the difficulty of the training task was incrementally increased over the three training blocks. This stepwise nature was structured in line with the number of perceptual features that needed to be considered to make the identification. In the first block, collages were graded as 1 (Normal) or more than a 1. This judgment could be made using only the first feature (Background). In the second block, collages were graded as a 1, 2, 3, 4, or more than 4. This judgment required participants to use the first two features (Background and Lines). In the final training block, collages were graded on the full 7-point scale. This judgment required using all three features (Background, Lines, and Gradient).

Table 1: The verbal descriptions included in the annotations in Experiment 1 for the three features (Background, Lines, and Gradient). The same description could apply to more than one grade, as indicated by "As above".

Grade	Background	Lines	Gradient
1	The liver tissue in the background (i.e., not directly adjacent to vessels) is not particularly bright.	Blood vessels have bright white lines adjacent to their walls.	The brightness of the lower tissue is similar to the upper tissue.
2	The liver tissue in the background (i.e., not directly adjacent to vessels) is brighter than normal (grade 1).	As above	As above
3	As above	Blood vessels have white lines adjacent to their walls, but these are generally less bright than in grade 2.	As above
4	As above	Some blood vessels do not have white lines adjacent to their walls. Some vessels do have white lines adjacent to their walls.	As above
5	As above	Most blood vessels do not have white lines adjacent to their walls.	As above
6	As above	As above	The lower tissue is slightly darker than the upper tissue.
7	As above	Almost none of the blood vessels have white lines adjacent to their walls.	The lower tissue is clearly darker than the upper tissue.

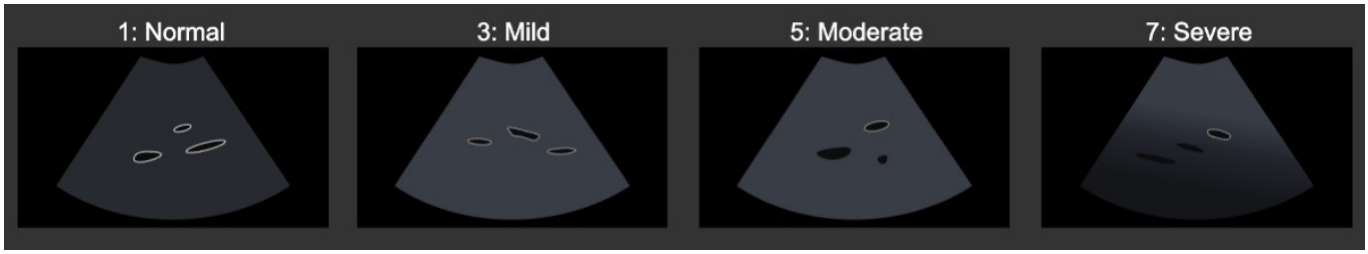

Figure 2: The four liver images shown to participants in the instructions at the start of Experiment 1 (depicting grades 1, 3, 5, and 7). Only a single liver image (instead of the entire collage) was shown.

Results

Participants took an average of 22 minutes to complete the experiment. Due to a technical error, 1-4 trials of data were missing for four participants, so analyses were conducted on their remaining data.

Figure 3 shows the mean error on the pre-test and post-test for each condition. It is evident that in all conditions, people were more accurate in the post-test (lower mean error). Collapsing across conditions, a paired-samples t-test found that the reduction in mean error from pre-test to post-test was significant, $t(199) = -18.77, p < .001$, 95% CI [-1.01, -0.81], $d = -1.33$. Therefore, we proceeded with our main analysis and conducted a 2 (Annotations) x 2 (Steps) between-participants ANOVA, with the mean difference in error between the pre-test and the post-test as the dependent measure. Contrary to our expectations, there was no significant main effect of ANNOTATIONS, $F(1,196) = 0.83, p = .364$. There was a significant main effect of STEPS, $F(1,196) = 6.51, p = .012, \eta^2 = .02$, although this was in the opposite direction than expected, with less improvement from pre-test to post-test ($M = -0.81$) for participants that underwent training in a stepped fashion compared to those that did not ($M = -1.03$). There was no significant interaction between ANNOTATIONS and STEPS, $F(1,196) = 1.88, p = .172$.

Discussion

We assessed whether ANNOTATIONS and STEPS provided benefits to learning in a task with artificial liver ultrasound images. Contrary to our predictions and prior studies (Frank et al., 2020; Johnston et al., 2020; Miyatsu et al., 2019), we did not find a benefit to supplementing perceptual training with annotated feedback, even though the categorization was perfectly determined by the perceptual features (i.e., no noise). One possibility for this finding is that the annotations were ineffective in our task because they were not intuitively obvious and required task-specific knowledge to be understood. Although Miyatsu et al. (2019)'s paradigm involved feature highlighting that related to multiple features, the na-

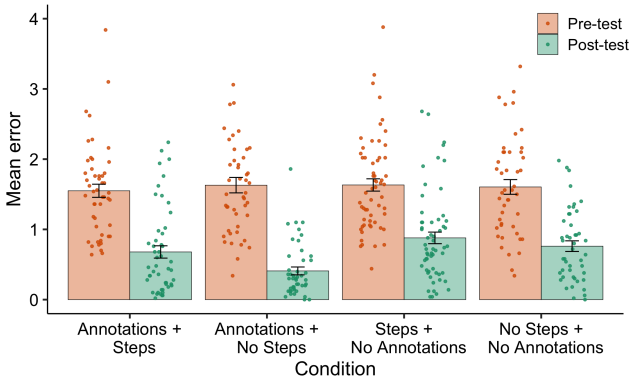

Figure 3: Performance by training condition on the pre-test and post-test. The *y* axis shows the mean error (distance from the correct grade); thus, lower is better. Each dot is the mean error for one individual, and error bars represent the standard error. All training conditions show improvement from the pre-test to the post-test.

ture of their stimuli (rocks) likely meant the features were easier to describe verbally (e.g., "grey to white crystalline material" and "darker swirls and veins"). Conversely, in our task, providing a cue that "the background liver tissue is darker than Normal", may not be helpful if the participant has not gained enough experience with how dark normal cases are. Additionally, the extent to which the visual manifestation of the features is similar between conditions (e.g., the brightness level of the walls around the blood vessels) likely increases the task difficulty and decreases the diagnostic value of the cues.

Inconsistent with our expectations, there was a negative effect of STEPS. One explanation for this is that in the stepped condition participants gain less experience with training on the task they were assessed on in the post-test. In particular, only in a third of the training trials were participants using all three cues and rating stimuli using the full 7-point range. An alternative possibility for why the stepwise training was not more successful is because the task required attention to disparate locations across the entire collage. This is unlike the previous studies in the perceptual training literature with medical images, which tend to focus on tasks that involve searching for a single target (e.g., identifying whether a tumour is present or not), and attending to a specific location in a single image. It could be that attending to multiple locations and multiple cues introduces a cost that cannot be surmounted by perceptual training.

Experiment 2

The aim of our second experiment was to investigate the potential cost of needing to attend to multiple locations and cues during learning. To better focus on this issue, we simplified the task by only manipulating a single feature (brightness) and by using simpler stimuli (disks). When the feature was split across multiple locations, the brightness levels followed a similar structure to the feature rules in Experiment 1. Therefore, in those conditions, the task could be completed in a similar stepwise fashion, such that in some cases the brightness of one location was sufficient to diagnose the grade, whilst in other cases, the brightness of two or all three locations was needed to identify the grade. This experiment allows us to gain insight into the extent to which splitting the cues across disparate locations increases the task difficulty. Additionally, we explored to what degree featural discriminability (i.e., how easy it is to discriminate between the different brightness levels at a single location) impacted learning.

Based on our findings in Experiment 1, we hypothesised that a task that requires attention to multiple locations will be more difficult than a task that only requires attention to a single location. Additionally, we expected that when the brightness levels at a single location corresponding to different grades are more difficult to discriminate (i.e., more similar to each other), performance will be negatively impacted.

Method

Participants We recruited 155 participants on MTurk with the same inclusion criteria as in Experiment 1, except all resided in the USA. Participants were compensated $4.50, with a $1 bonus awarded to the top 20% of performers. Consistent with our pre-registration `https://aspredicted.org/VBY_HBY`, one participant was excluded for incomplete data and one for failing attention checks. The final dataset included 153 people (75 males; one unreported) with a mean age of 40.4 years ($SD = 11.0$).

Design and Procedure Participants were randomly assigned to one of three conditions. Figure 4 provides the category structure and an example stimulus for each condition. In CONDITION 1, the stimulus was a single uniform disk that could vary in brightness across one of seven shades of grey. Each brightness level corresponded to a different grade on the same 7-point scale that was used in Experiment 1. In CONDITION 2, the stimulus consisted of three disks that could each take on one of three levels of brightness, corresponding to the lowest, middle, and highest brightness levels in CONDITION 1). CONDITION 3 was similar to CONDITION 2, except that for each disk the brightness range of the three levels was equal to approximately one-third of the range of the brightness levels in CONDITION 1. Thus, the total range summed across all three locations in CONDITION 3 equalled the range of brightnesses in CONDITION 1. In CONDITION 2 and CONDITION 3, the mapping between the grade and the brightness of the second disk was reverse coded in order to prevent judgments being made on the overall brightness of the three disks.

The experiment was self-paced and completed online. Participants were informed that they would need to grade each stimulus according to a 7-point grading scale and were provided with four examples of the stimuli (grades 1, 3, 5, and 7) prior to completing a pre-test (50 trials). Following this, participants underwent a training phase (90 trials) and then had their performance assessed in a post-test (50 trials). In each phase of the experiment, the stimuli were graded according

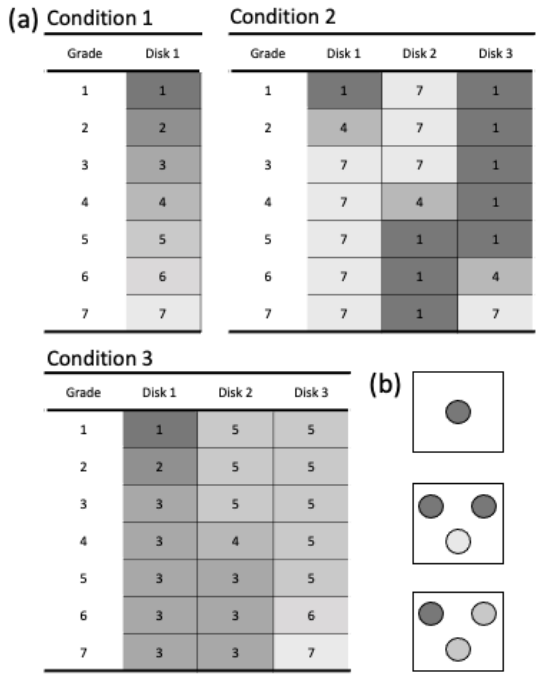

Figure 4: (a) The stimuli structure used in Experiment 2 for each condition. The numbers in the disk columns depict the brightness level, with the cell fill depicting the shade that was used. (b) An example stimulus for each condition (1, 2, and 3, respectively) in Experiment 2. Each of the examples depicts a grade 1 (Normal).

to the same 7-point scale. No feedback was provided during the test phases. During the training phase, participants were provided with corrective feedback that informed them of the correct grade.

Results

Participants took an average of 21 minutes to complete the experiment. As one trial of data was missing for one participant due to a technical error, analyses were conducted on their remaining data.

Figure 5 shows that the mean error was lower on the post-test than the pre-test for all conditions, suggesting that some learning occurred. Additionally, as shown by the individual data points in CONDITION 1 (single-disk), the pre-test data was bimodal. Two participants from the cluster of higher mean error in CONDITION 1 reported that they had mistakenly reversed their use of the scale throughout the pre-test (e.g., grading stimuli with darker shades of grey as more severe, contrary to the example stimuli shown in the instructions). Additionally, as can be seen in Figure 6, the mean error during the training phase rapidly decreased at the start of CONDITION 1. This supports the possibility that some participants may have been using the scale in reverse during the pre-test and then subsequently corrected their misunderstanding once they received feedback in the training phase. Therefore, to reduce the possibility of this biasing our results, we chose to conduct our analysis on the post-test data only, instead of on the mean difference between the pre-test and

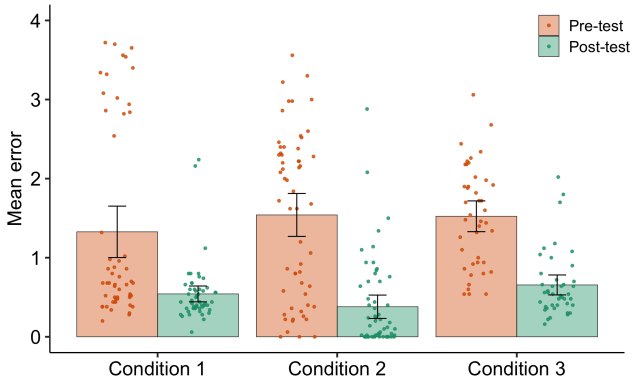

Figure 5: Performance by training condition on the pre-test and post-test. The *y* axis shows the mean error (distance from the correct answer); thus, lower is better. Each dot is the mean error for one individual, and error bars represent the standard error. All training conditions show improvement from the pre-test to the post-test.

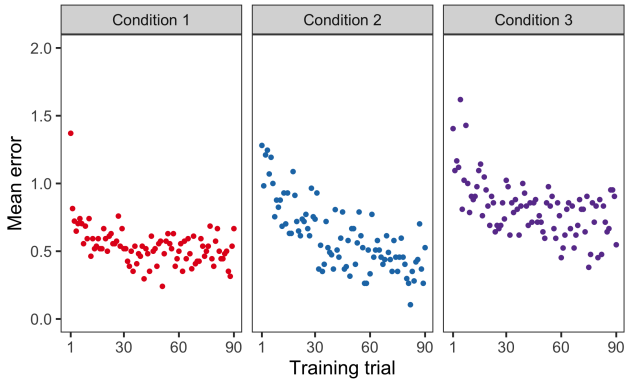

Figure 6: Mean error over the course of the training for each condition, where the *x* axis shows the training trial number. There is a downward trend in mean error across all conditions over the course of the training. The mean error in CONDITION 1 (red dots) rapidly decreases from the first trial.

post-test (which was the analysis we had pre-registered).

As assumption checks indicated the assumption of normality was violated, instead of a one-way ANOVA, we conducted a Kruskall-Wallis test, and found a significant difference in the mean post-test error between conditions, $\chi^2(2) = 23.90, p < .001$. Post hoc pairwise comparisons using Dunn's test with a Bonferroni correction found that CONDITION 2 (multi-disk, full brightness range for each disk) had significantly lower mean error ($M = 0.38; SD = 0.57$) than CONDITION 1 (single-disk; $M = 0.54; SD = 0.38$) and CONDITION 3 (multi-disk, reduced brightness range for each disk; $M = 0.66; SD = 0.42$), $p < .001$. The difference between CONDITION 1 and CONDITION 3 was non-significant, $p = .477$.

Discussion

Our predictions were not supported, as we found that attending to multiple locations did not negatively impact learning, compared to attending to a single location. Our findings suggest that learning may be impacted most by how discriminable the feature levels are from each other. Specifically,

when the brightness levels are spread out further as in CONDITION 2, the task is easier to learn because the brightness levels are easier to discriminate.

General Discussion

In our first experiment, we developed a noise-free artificial stimulus set, based on an actual medical image data set, to rigorously investigate the effectiveness of augmenting perceptual training with annotations and stepped instruction in a 7-point discrimination task. We found that for a task that requires attention to multiple features that are difficult to verbally describe, annotations or a stepwise training paradigm did not provide additional benefits to learning, beyond a standard perceptual training approach.

Perhaps this is not so surprising as feature highlighting has been found to be most beneficial when the features are easily interpretable by the learner (Meagher et al., 2022). Conversely, in Experiment 1, understanding the significance of the features required task-specific knowledge, such as the typical brightness level for normal liver tissue. It could be that our participants lacked the task-specific knowledge they needed to utilise the annotations.

It is possible that the stepped learning procedure was not more beneficial because it reduced the number of training trials where participants practiced on the task that they were subsequently tested on. In the post-training test phase, participants were required to utilise all three features and to distinguish all 7 levels of hepatic steatosis. In the stepped training, they practice doing this on only a third of the trials. Conversely, in the standard training, they practiced doing this on all the training trials.

The second experiment investigated the potential cost of attending to multiple locations and whether performance was affected by the discriminability of the different brightness levels at each location. As expected, performance was highest when it was easiest to discriminate the different brightness levels. However, contrary to our expectations, there was no decrement in performance when participants were required to attend to three locations as opposed to just one location.

From the above, it follows that the task in Experiment 1 was likely intrinsically difficult not because the participants needed to attend to multiple locations but because it was difficult to discriminate the different levels of each feature. For example, the first feature (Background) had just two brightness levels, normal and brighter than normal. Participants may have had difficulty using this feature to discriminate between a grade of 1, as indicated by the background having a normal level of brightness, from a grade greater than 1, as indicated by the background having a brighter than normal level of brightness, due to the difficulty of distinguishing between these two brightness levels. Similarly, participants may have had difficulty distinguishing between the five levels of the second feature (Lines) and the three levels of the third feature (Gradient). If so, this would have made it hard for them to discriminate grades 2-7.

One might have expected that perceptual training would have improved our participants' ability to distinguish between these different feature levels. Indeed, previous studies have shown that perceptual training can lead to participants being able to make finer perceptual discriminations (Sagi, 2011). Why then did we not observe a greater improvement in performance?

It is possible that the training phase in our experiments was too short. Our experiments used a relatively short amount of training (90 trials), whilst prior perceptual training studies have tended to involve larger amounts of training (e.g., 100s or even 1000s of trials). For example, Chen et al. (2017) found that after 1280 training trials, novices were able to achieve approximately the same level of performance as experts (radiologists) in a hip fracture identification task. It is possible that had the training phase in our experiments been longer, the performance of our participants would have improved further. That said, some previous studies (e.g. Marris et al., 2023) found little improvement beyond 90 trials on a similar task, so longer training may not change these results.

Our approach to testing these training paradigms with artificial stimuli provides some insight into the extent to which perceptual training paradigms may be useful with real images. As the artificial stimuli we used were highly controlled, it is expected that the effect of perceptual training with real liver ultrasound images (which contain noise) would be more limited and that learning would likely, therefore, be more gradual. We recommend that real-world perceptual training is designed to match the task of interest as closely as possible (stepwise training was detrimental to learning), and that standard perceptual training procedures are used, as annotations provided no substantial benefits.

In summary, our findings demonstrate that perceptual training can be extended to train people to perform a difficult perceptual discrimination task, although the extent of improvement is limited. Perceptual training may be a useful supplement to existing training regimes in the medical domain but is not a replacement for the existing training that professionals currently receive.

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

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. doi: 10.1371/journal.pone.0189192

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12. doi: 10.3758/s13428-014-0458-y

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. doi: 10.1016/j.cub.2020.05.050

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. doi: 10.1167/jov.20.8.16

Marris, J., Perfors, A., Mitchell, D., Wang, W., McCusker, M. W., Lovell, T. J. H., ... Howe, P. D. L. (2023). Evaluating the effectiveness of different perceptual training methods in a difficult visual discrimination task with ultrasound images. *Cognitive Research: Principles and Implications*, *8*, 1-18. doi: 10.1186/s41235-023-00467-0

Meagher, B. J., McDaniel, M. A., & Nosofsky, R. M. (2022). Effects of feature highlighting and causal explanations on category learning in a natural-science domain. *Journal of Experimental Psychology: Applied*, *28*(2), 283–313. doi: 10.1037/XAP0000369

Miyatsu, T., Gouravajhala, R., Nosofsky, R. M., & McDaniel, M. A. (2019). Feature highlighting enhances learning of a complex natural-science category. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(1), 1–16. doi: 10.1037/XLM0000538

Sagi, D. (2011). Review: Perceptual learning in Vision Research. *Vision Research*, *51*, 1552–1566. doi: 10.1016/j.visres.2010.10.019

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), 4. doi: 10.1186/s41235-020-0208-x

Van Merriënboer, J. J., Kirschner, P. A., & Kester, L. (2003). Taking the load off a learner's mind: Instructional design for complex learning. *Educational Psychologist*, *38*(1), 5–13. doi: 10.1207/S15326985EP3801_2