Perceptual Learning of Appendicitis Diagnosis in Radiological Images

Ian Andrew Johnston Department of Psychology, University of Wisconsin-Madison, Madison, WI, USA

Mohan Ji Department of Psychology, University of Wisconsin-Madison, Madison, WI, USA

Aaron Cochrane Department of Psychology, University of Wisconsin-Madison, Madison, WI, USA

Zachary Demko Department of Psychology, University of Wisconsin-Madison, Madison, WI, USA

Jessica B. Robbins Department of Radiology, University of Wisconsin School of Medicine and Public Health, Madison, WI, USA

Jason W. Stephenson Department of Radiology, University of Wisconsin School of Medicine and Public Health, Madison, WI, USA

C. Shawn Green Department of Psychology, University of Wisconsin-Madison, Madison, WI, USA

A sizeable body of work has demonstrated that participants have the capacity to show substantial increases in performance on perceptual tasks given appropriate practice. This has resulted in significant interest in the use of such perceptual learning techniques to positively impact performance in real-world domains where the extraction of perceptual information in the service of guiding decisions is at a premium. Radiological training is one clear example of such a domain. Here we examine a number of basic science questions related to the use of perceptual learning techniques in the context of a radiology-inspired task. On each trial of this task, participants were presented with a single axial slice from a CT image of the abdomen. They were then asked to indicate whether or not the image was consistent with appendicitis. We first demonstrate that, although the task differs in many ways from standard radiological practice, it nonetheless makes use of expert knowledge, as trained radiologists who underwent the task showed high (near ceiling) levels of performance. Then, in a series of four studies we show that (1) performance on this task does improve significantly over a reasonably short period of training (on the scale of a few hours); (2) the learning transfers to previously unseen images and to untrained image orientations; (3) purely correct/incorrect feedback produces weak learning compared to more informative feedback where the spatial position of the appendix is indicated in each image; and (4) there was little benefit seen from purposefully structuring the learning experience by starting with easier images and then moving on to more difficulty images (as compared to simply presenting all images in a random order). The implications for these various findings with respect to the use of perceptual learning techniques as part of radiological training are then discussed.

Introduction

Humans are excellent at perceptual learning. Given appropriate experience, humans will tend to drastically improve their ability to extract perceptual information from the environment and to make decisions based

Citation: 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):16, 1–17, https://doi.org/10.1167/jov.20.8.16.

https://doi.org/10.1167/jov.20.8.16 Received April 2, 2020; published August 13, 2020 ISSN 1534-7362

upon the extracted information (Dosher & Lu, 2017; Green, Banai, Lu, & Bavelier, 2018; Sagi, 2011; Seitz, 2017). Indeed, the visual perceptual learning literature offers many examples of dedicated training resulting in marked enhancements in participants' ability to make exceptionally fine perceptual judgments. Such improvements have been demonstrated for a wide variety of visual stimulus dimensions, including stimulus orientation (Vogels & Orban, 1986), relative spatial position (McKee & Westheimer, 1978), contrast (Sowden, Rose, & Davies, 2002), spatial frequency (Fiorentini & Berardi, 1980), motion direction (Ball & Sekuler, 1987), and motion speed (Saffell & Matthews, 2003). Yet, despite this extensive body of work demonstrating the utility of training paradigms designed to enhance humans' ability to make decisions about single well-defined dimensions of reasonably simple sensory stimuli (e.g., decisions about the orientation of black-and-white gratings or the direction of moving dots), there remain a number of open questions regarding how to best extrapolate from the existing body of perceptual learning knowledge so as to enhance performance in various real-world domains. This is particularly true in the case of real-world domains where the extraction of decision-relevant information is similarly at a premium, but where the stimuli, the perceptual dimensions of interest, and the decisions are more complex than what has typically been considered in the perceptual learning literature to date.

One potentially significant obstacle standing in the way of translation from the basic science perceptual learning literature to more real-world domains is that the improvements that have been documented to emerge through typical perceptual training are commonly quite specific to the exact details of the trained stimuli. For example, individuals who have learned to make extremely subtle judgments about the direction in which a pair of vertically oriented lines are offset from one another may subsequently show no improvements in their ability to make seemingly comparable offset judgments about horizontally oriented lines (Fahle, 1997). Similarly, individuals who have increased their ability to discriminate certain visual targets that have always been presented in one particular spatial location throughout training may then show no enhancements in their ability to discriminate the exact same visual targets when they are presented in untrained spatial locations (Zhang & Li, 2010). This tendency has sometimes been referred to as the "curse of specificity" in the context of real-world applications, with the "curse" referring to the fact that nearly all real-world tasks necessitate reasonable generality of function (Deveau & Seitz, 2014; Green & Bavelier, 2008). After all, in the real world, stimuli can easily appear in new and somewhat unexpected locations, at new orientations, or in different sizes (particularly retinal sizes, which will depend on viewing distance). Therefore, training that produces benefits for stimuli of a single orientation, size, spatial frequency, and retinal location is likely to have limited real-world value. One major question to be addressed, then, is whether such learning specificity will be as prevalent given training on the complex real-world stimuli relevant to the given domains.

A second major question that has emerged with respect to the use of perceptual learning techniques in more real-world contexts revolves around the importance of feedback during training (Dosher & Lu, 2009; Herzog & Fahle, 1997; Liu, Lu, & Dosher, 2010; Petrov, Dosher, & Lu, 2006; Seitz, Nanez Sr., Holloway, Tsushima, & Watanabe, 2006). Although it is widely recognized that the presence of informative feedback can be beneficial for perceptual learning, in practice there are many examples of individuals learning to perform perceptual learning-type tasks even in the absence of informative feedback (Fahle, Edelman, & Poggio, 1995; Karni & Sagi, 1991). The extent to which informative feedback is necessary in various real-world tasks is largely unknown; yet, this is a critical issue. In many real-world domains, the natural learning experience does not necessarily always come with immediate informative feedback.

Finally, a last question regarding translation of perceptual learning techniques into more real-world domains centers on how to best structure task difficulty throughout training so as to maximize both learning of the task itself and to maximize the extent to which learning generalizes to new stimuli or contexts (noting from the offset that these two goals may be mutually exclusive in some situations; see Bavelier, Bediou, & Green, 2018). In general, there is wide-ranging agreement that task difficulty during training impacts the final learning outcome. In fact, this broad phenomenon is not unique to the perceptual learning literature, but instead is seen throughout the much wider learning domain that encompasses everything from motor learning to learning in educational context (Guadagnoli & Lee, 2004; Vygotsky, 1978). However, universal best practices remain elusive. In particular, it is unknown whether the same outcomes related to structuring task difficulty during training will be observed in perceptual learning tasks where key features lie along multiple perceptual dimensions rather than a single dimension (e.g., orientation, spatial frequency).

Here, we consider these issues in the context of a real-world domain that, at least on the surface, appears to be well suited for the use of perceptual learning techniques to enhance current practice—learning to make radiological diagnoses (Kellman, 2013; Kelly, Rainford, McEntee, & Kavanagh, 2018; Kundel & Nodine, 1983; Li, Toh, Remington, & Jiang, 2020; Sowden, Davies, & Roling, 2000; Waite et al., 2020).

First and foremost, radiological practice clearly involves learning to extract relevant perceptual information in the service of effective decision-making (Kelly, Rainford, Darcy, Kavanagh, & Toomey, 2016; Ravesloot et al., 2017). Throughout a single day, a practicing radiologist will view a substantial number of cases that may include x-rays, computed tomography (CT) scans, and/or magnetic resonance imaging. And, in each case, they are asked to make a series of decisions about what the scans indicate. Second, existing training for radiologists is potentially inefficient, given the goal of learning to make effective perceptual decisions. Radiologists, like most medical professionals, are currently primarily trained using an apprenticeship model. Trainees predominantly see the radiological cases that happen to come into the medical facility during the time period when they are working. As such, almost by definition, the experience is largely unstructured (e.g., in terms of what diagnoses might be encountered, in what order, with what degree of variety, with what idiosyncrasies). Furthermore, feedback can be significantly delayed relative to when the decisions are actually made. This is a significant issue, as a recent study by Li et al. (2020) found that accurate feedback was critical in learning about abnormalities in radiological images.

In addition to being markedly more efficient in terms of the total number of cases seen per unit time and in terms of training structure, the use of perceptual learning techniques could also allow for the inclusion of difficult or rare cases that are almost never seen in practice. This would ensure that the first time an individual sees such an exemplar it is not in the context of a serious medical issue being experienced by a real-world person. Furthermore, training can be repeated or augmented as needed, and training may provide an objective measure of one's ability to make the necessary diagnoses. Each of these features would be significant improvements over the current apprenticeship model.

The global aim of this study was thus to begin to explore many of the issues identified above in the context of a radiological task. In particular, our key questions included the following: (1) Do perceptual learning techniques allow naïve participants to learn to accurately make a common diagnosis of appendicitis in a short timeframe? and (2) If learning does occur, will it transfer to new contexts (i.e., new images, new orientations)? Given positive results for questions 1 and 2, we then moved on to the more secondary questions: (3) Is informative feedback necessary for learning effects to be observed? and (4) Does deliberate scaffolding of difficulty enhance learning outcomes? Before considering these key questions via training of naïve individuals, though, we first examined whether our task made appropriate use of radiological knowledge.

Study 1: Do radiologists show expert performance on the to-be-trained stimuli?

Given the relatively basic-science, perceptual-learning-related questions of interest in the current work, single representative CT images were preferred for training and testing of novices. Yet, our goal was, at the same time, to construct a task that, although being more experimentally tractable than real-world situations, nonetheless relied on the same core perceptual expertise that develops during radiological training. In other words, because single representative CT abdominal images are clearly simplified relative to actual radiological practice (i.e., because CT scans normally allow a radiologist to traverse a full series of two-dimensional images that progress through the body, creating a three-dimensional volume), we first sought to ensure that the single images still provided a good model of developed radiological expertise. Thus, the goal of Study 1 was to assess whether trained radiologists showed expert performance when presented with these single-image radiological stimuli. If this was indeed the case, we would then feel more confident moving forward with these stimuli for training and testing (in Studies 2–5).

Methods

Participants

The expert cohort was composed of 12 practicing radiologists, all from the University of Wisconsin-Madison. Four of the participants were abdominal imaging fellows (5 years of direct experience in the field of radiology), and eight participants were fellowship-trained abdominal imaging attending physicians (average of 18 years of experience in radiology; range, 6–33 years).

Stimuli

Two hundred unique cases were randomly identified from clinical cases found in a University of Wisconsin picture archiving and computing system (PACS): 100 cases with a normal appendix and 100 cases of acute uncomplicated appendicitis. An expert radiologist (authors J.R. or J.S.) selected the single best axial image depicting the appendix for each case (Figure 1). These single images were saved as anonymous JPEG files captured from the original Digital Imaging and Communication in Medicine (DICOM) de-identified clinical images at a resolution of 512 × 512 pixels. No images or image file names included identifiable personal health information. Static images were

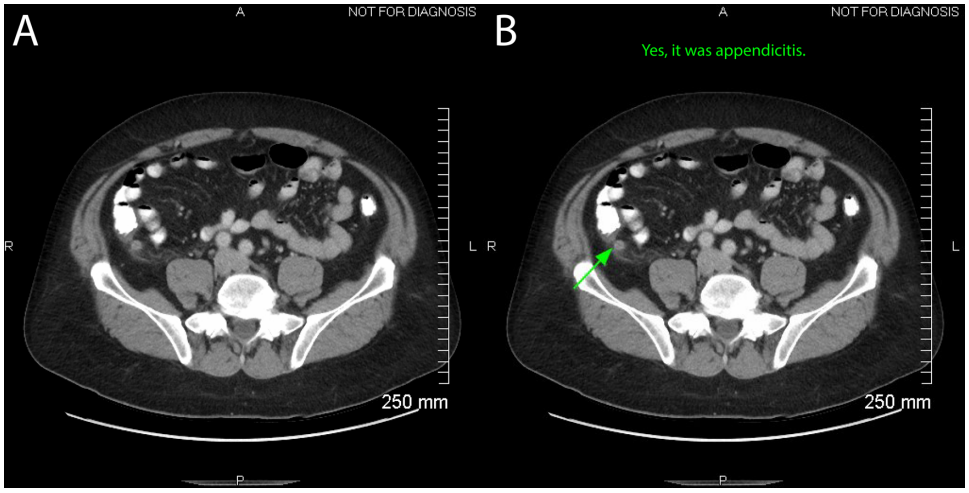

Figure 1. Stimulus and feedback screens. (A) On each trial, participants were presented with a single axial image from a CT scan of the abdomen (image chosen to best depict the appendix). They were asked to indicate whether they believed the image was or was not consistent with appendicitis via a button press. (B) In those studies that included feedback (Studies 2–5), following their response, participants were shown the same axial image with text indicating whether their response was correct or not (on the top of the screen). In Studies 2, 3, and 5 an additional form of feedback was also presented in the form of a green arrow pointing at the location of the appendix on the image.

displayed to the participants; the participants were unable to window or level the images. Two additional sets of images, referred to below as the "flipped" images, were created by inverting a subset of images from the full set above either across the vertical axis (referred to as "left–right flipped"; 40 images) or across the horizontal axis (referred to as "up–down flipped; 40 images). There was no overlap between the sets, and the percentages of normal appendix and appendicitis images were balanced within each set. The same sets were used across all participants.

Procedure

The task was presented on a Windows laptop with a 17-inch display (Microsoft, Redmond, WA) in a quiet room in the radiology facilities at the University of Wisconsin hospital. On each trial of the task, participants viewed one of the single axial CT images. The images remained present until the participant responded as to whether or not he or she believed the image was consistent with appendicitis by pressing one of two buttons. No feedback was given after responses. Participants first completed 200 trials with unflipped images (i.e., each of the possible images presented once in pseudorandom order) followed by 80 trials with the flipped images (i.e., each of the possible flipped images presented in pseudorandom order).

Data analysis

Results in all of the studies below rely on comparisons or assessments of fit parameters from nonlinear regressions that allow for possible monotonic changes in d' over the course of each relevant block of trials. Our previous work has shown that modeling possible continuous trial-by-trial improvements in perceptual learning tasks provides a number of systematic benefits in terms of quality of fit and the inferences that can be made over coarser aggregate measures, such as simply calculating d' over the entirety of sessions or blocks (Kattner, Cochrane, Cox, Gorman, & Green, 2017; Kattner, Cochrane, & Green, 2017). The nonlinear regressions utilized here parameterize performance as a starting level, an asymptotic level, and a rate of change between start and asymptote. Rate is defined as a binary log time constant (i.e., smaller rate values indicate faster change). Nonlinear regressions were fit using the R package TEfits (R Foundation for Statistical Computing, Vienna, Austria) (Cochrane, 2020). Specifically, within TEfits, smoothed proportions of false alarms (pFA) and smoothed proportions of hits (pH) were transformed to d' and then fit, using maximum-likelihood optimization, to each participant's data using nonlinear least squares and the above-described three-parameter exponential function (start, asymptote, and rate). pFA and pH were independently smoothed using a Gaussian kernel with a half width at half maximum of five trials. We calculated the cumulative Gaussian density at –2.5, then applied this as an edge offset (e.g., approximately $0.0062 + [1–2 \times 0.0062] \times \text{pFA}$).

TEfits was used to pass these initial maximum-likelihood fits to Bayesian model fitting using the R package brms. Following model fitting and convergence checks in brms (all R-hat < 1.05), by-participant

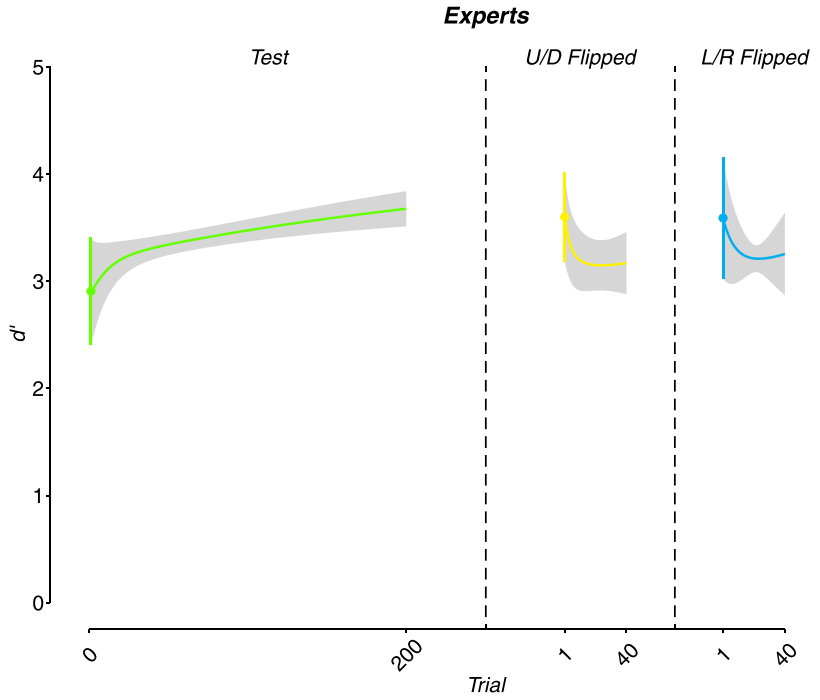

Figure 2. Expert (trained radiologist) performance. As expected, trained radiologists showed extremely high levels of performance on the task from the outset. Although there was a slight trend toward improved performance as a function of trial number, this was not reliable. There was also no reliable difference in performance between the test images and the images that were flipped either up–down (U/D) or left–right (L/R). Note that, in this and all further figures of this type, colored lines going in the horizontal direction depict the fit line. Colored vertical lines and shade areas represent the standard error of the mean (*SEM*).

parameters were used in subsequent analyses. By-participant point estimates were utilized for flexibility in within-study as well as cross-study comparisons and for evaluating performance at certain trial numbers. We note here that, although it would be possible to derive any particular hypothesis test directly from the fit Bayesian hierarchical models, we instead chose to use by-participant point estimates so that results from each experiment would be directly comparable in terms of d'. In addition to standard significance values, we also report the Bayes factor (*BF*) associated with tests when applicable. Bayes factors are reported on a log 10 scale to assist with interpretability; for example, a *BF* of 0.3 indicates that there is about twice as much evidence for the alternative than the null, a *BF* of –0.6 indicates that there is about four times as much evidence for the null than the alternative, and a BF of 1 indicates 10 times as much evidence for the alternative than the null.

Results

As expected, the experts showed a high degree of proficiency on the task (note that in terms of raw accuracy these values universally correspond to >90% performance). The experts' estimated starting performance was significantly different from 0: $M = 2.91 \pm 0.50$ (for this and further reported statistics, error represents the standard error of the mean), $t(11) = 5.82$, $p < .001$, and $BF = 2.40$. Although we did observe a slight numerical increase in performance throughout the duration of the experiment in the normal set of images (Figure 2), this increase was not significant; by contrasting the estimated starting performance with the estimated final performance, $t(11) = 1.42$, $p = 0.18$, and $BF = 0.19$. The experts also showed no significant difference in the estimated starting performance for either type of flipped image sets as compared to the normal image set: for up–down flipped, $M = 3.6 \pm 0.42$, $t(11) = 1.04$, $p = 0.319$, and $BF = –0.34$; for left–right flipped, $M = 3.59 \pm 0.57$; $t(11) = 1.09$, $p = 0.297$, and $BF = –0.33$.

Discussion

The results of Study 1 demonstrate that, despite the simplified nature of the stimuli, they nonetheless were sufficient for expert radiologists to make an appropriate appendicitis/normal appendix diagnosis with a high degree of accuracy. This in turn suggests that the task is tapping at least some of the perceptual expertise that develops through radiological training. Further, the experts performed similarly on the normal images and

the flipped images (both left–right flipped and up–down flipped). This could reflect a variety of possible causes, including learning generalization (i.e., from their trained perceptual expertise to the flipped images) or actual experience (e.g., reading scans in a variety of orientations depending on how the patient was positioned during the given scans). We would further note that because we utilized a "normal images → flipped images" design (to be more analogous with the "train → transfer" design used in the studies below), we do not have an estimate of performance on the flipped images taken before the normal images. Given this, it is also possible that if performance on the flipped images was measured prior to experience with the normal images it would have been poor. We can therefore not rule out the possibility that the equivalent performance on the flipped images as measured after experience with the normal images was a learning transfer effect. Although this is a potentially interesting question for future work, it does not bear on our main questions of interest. We thus moved forward in Study 2 to address whether or not novice individuals could learn to make appropriate diagnoses given training.

Study 2: Are naïve participants capable of learning to differentiate appendicitis/no appendicitis in a relatively short period of perceptual training?

The results of Study 1 indicated that the to-be-utilized stimuli and task rely upon the same type of perceptual expertise as developed during radiological training; thus, we then moved to our primary basic science research questions.

As an initial exploration, we chose to examine whether reasonably standard perceptual learning techniques (i.e., two-alternative forced choice) would produce enhanced performance with the given stimuli and task. More specifically, we had two key initial questions. Our first question was whether the task was, in fact, learnable given an amount of experience that would be typical in perceptual learning designs. The presence of learning is a necessary condition to then be able to ask more detailed follow-up questions regarding learning specificity or regarding training task structure. Previous work has shown improved performance in naïve individuals given reasonably standard training in the context of femur fractures (Chen, HolcDorf, McCusker, Gaillard, & Howe, 2017). However, it seems likely that the visual features indicative of such fractures are more obvious than is true of the features indicative of appendicitis. Second, given the somewhat limited set of total images (200 total), it was necessary to repeat images throughout training in order to have a sufficient total number of trials. We thus wanted to ensure that any learning that emerged was not at the level of individual exemplar cases alone (i.e., that learning, at a minimum, generalized to previously unseen images).

We chose, in Study 2, to first investigate the questions above using a sample of reasonably trained psychophysical observers from the lab of the senior author (C.S.G.) rather than fully naïve participants. This methodological approach of using trained psychophysical observers has a long history in the perceptual learning literature, particularly in situations where the main question pertains to whether a given skill can be improved through practice (De Valois, 1977; Poggio, Fahle, & Edelman, 1992; Cochrane, Cui, Hubbard, & Green, 2019). The general reasoning underlying this choice is that, if the skill is not seen to be learnable in a more select population, then it is unlikely to be learnable in a less select set of individuals. If the task was shown to be learnable in this more selected set of participants, we could then move forward testing participants who were not trained psychophysical observers (Studies 3–5).

Methods

Participants

Eight participants (mean age, 20.8 years; three females) took part in Study 2. All participants had experience with perceptual learning tasks by nature of their belonging to the lab of the senior author (C.S.G.). One of the participants was author A.C. Note, however, that A.C. only came to be part of the experimenter/author team after taking part as a participant in Study 2. Despite their general psychophysical task experience, no participants had medical training, prior exposure to CT image interpretation, an understanding of the to-be-trained stimuli (e.g., that there were 200 total images, that there were to be flipped versions of the images), or knowledge of the goals or hypotheses at hand.

Stimuli

The same 200 unflipped images as in Study 1 were utilized. These unflipped images were divided into a training set of 160 images (80 normal appendix and 80 appendicitis) and a test set of 40 images (20 normal appendix and 20 appendicitis). As noted below, the training set images were seen a total of five times each during training, whereas the test set images were unseen during training. Dividing the images in this manner allowed us to determine whether any learning that was observed on the trained images was specific to

the trained exemplars or generalized to the broader diagnosis of appendicitis in previously unseen images.

Procedure

All sessions of the experiment took place in a dimly lit testing room. Stimuli were presented on a Dell OptiPlex 780 computer with a 23-inch flat screen monitor (Dell Technologies, Round Rock, TX). Prior to beginning training, participants first viewed a 10-minute timed slideshow that oriented the participants to the general location and appearance of a normal appendix and the typical CT findings of appendicitis. After completing this preparatory activity, participants began the training phase. On each trial of the training task, participants viewed one of the single axial CT images from the training set. The images remained present until the participant responded as to whether or not he or she believed the image was consistent with appendicitis by pressing one of two buttons. After making a response, participants received immediate feedback (Figure 1B). The feedback indicated the accuracy of the response with a text display ("yes, it was appendicitis" or "no, it was not appendicitis") accompanied by a high-pitched beep for a correct response and a low-pitched beep for an incorrect response. On this feedback screen, the participant was also shown an annotated version of the CT image they had just responded to, with an arrow identifying the location of the appendix.

The training phase consisted of five separate training sessions, each performed on a separate day. In each session, all 160 possible training images were displayed once in a pseudorandom order. During the sixth and final session of the task, the test set images (that had not been seen during the prior five training sessions) were randomly intermixed with the training set images. Each session took participants approximately 20 to 25 minutes to complete.

Results

Exclusions

One participant did not complete all sessions of the experiment, and one participant did not follow task instructions (strongly prioritizing speed over accuracy). Both participants were therefore excluded from further analysis, resulting in a total of six participants being included in the final analyses.

Data analysis

We employed the same basic fitting method as described under Study 1. The one difference for the current analysis was that for the training data the initial performance parameter was fixed at 0. This was done for both theoretical and practical reasons. In terms of theory, given that the participants had no previous experience with the task or radiological images more generally at the beginning of training, they were expected to perform at chance level on trial 1. In terms of practice, we found that the inclusion of a free parameter for the starting performance was not justified in terms of overall improvements in the quality of fit (particularly as the starting performance parameter is necessarily associated with more uncertainty than the learning rate or asymptotic performance parameters).

Learning of trained images

As expected, the participants showed clear evidence of learning, with their estimated final performance being significantly different from 0: $M = 1.59 \pm 0.21$, $t(5) = 7.69$, $p < 0.001$, and $BF = 1.78$ (Figure 3).

Generalization to test images

Given that participants showed clear evidence of learning on the training set of images, we then examined the extent to which the learning generalized to previously unseen images. We first compared participants' start performance on the test set of images to 0. This analysis tested whether the participants performed better on the test images following training than would be expected if they were fully naïve to the task; that is, a significant result would indicate that there was at least partial transfer of learning to the new images. The estimated starting performance on the test images following training was significantly greater than 0: $M = 2.07 \pm 0.65$, $t(5) = 3.18$, $p = 0.02$, and $BF = 0.55$ (Figure 3). We next compared the participants' performance on the test set of images to their estimated final performance on the training set of images. This analysis assessed whether the participants performed similarly on the test images and the trained images following training. A significant result here would indicate that transfer was not "full transfer" or, in other words, that there was some fall off in performance on the new images; a non-significant result would indicate a lack of evidence for such specificity of learning. The participants' estimated starting performance on the test set of images was not significantly different from the estimated final performance on the training set of images: $t(5) = -0.63$, $p = 0.554$, and $BF = -0.36$ (Figure 3). Participants' estimated final performance on the test set of images was also compared with the estimated final performance on the training set of images. Consistent with our other results, we did not find evidence of different levels of performance: $t(5) = -1.29$, $p = 0.252$, and $BF = -0.17$.

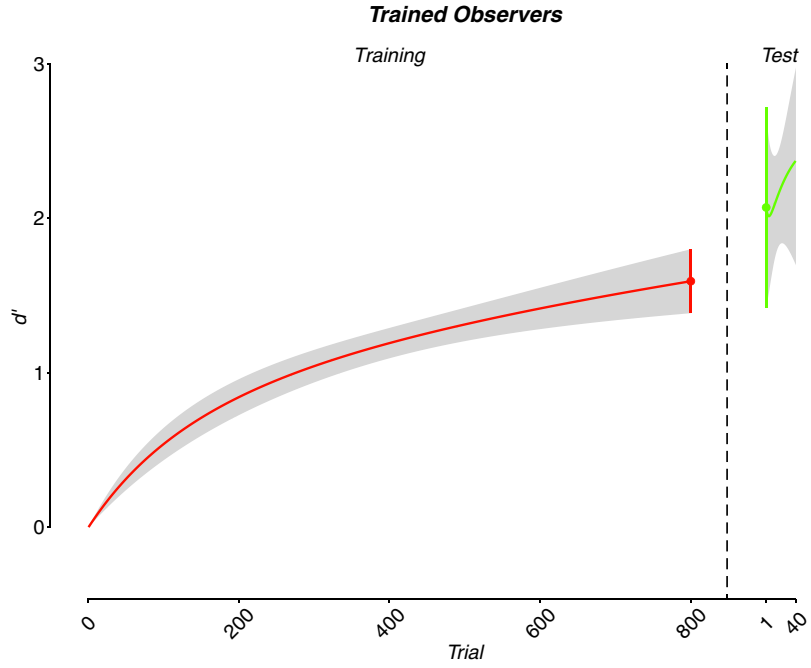

Figure 3. Trained psychophysical observers (although naïve to task). Participants showed clear evidence of learning through time. There was no significant difference between performance on the trained images at the conclusion of training and the test images, indicating that the learning was not specific to the images utilized during training but instead extended to previously unseen images.

Discussion

Overall, the results of Study 2 indicate that (1) the task was indeed learnable within a reasonably short time frame by individuals inexperienced with radiological images but who did have experience with other psychophysical learning tasks, and (2) the learning generalized to new images.

Study 3: Do fully naïve participants show a similar pattern of results in terms of learning and learning generalization?

Given the results of Study 2, we addressed two new questions in Study 3. Our first question was whether the same basic results seen in Study 2 would hold in a less selected group of participants (i.e., in naïve undergraduates rather than undergraduates with some degree of experience with psychophysical tasks). This is a key question given earlier work showing that previous experience with psychophysical tasks can improve the ability to learn new tasks that share similar structure. Our second question was, if learning does indeed occur, is it specific for the trained orientation/spatial position of the stimuli? Indeed, as noted earlier, one common outcome of perceptual training is that the learning that emerges is highly specific to certain aspects of the trained stimuli, with the orientation and spatial position of the stimuli being two of the more prominent dimensions, along with specificity of learning, as has been previously observed. To address this issue, at the conclusion of training, participants performed the basic task not only with the same test images as in Study 2 (i.e., to examine generalization to previously unseen images) but also with the "flipped" images (from Study 1).

Methods

Participants

Sixty-two participants enrolled in the study in exchange for course credit. Technological issues resulted in demographic data not being properly saved for 22 participants; the mean age of the remainder of the sample was 20.26 ± 2.67 years (20 females). The participants had limited or no experience with psychophysical tasks generally and no experience with radiological images of any sort.

Stimuli and procedure

The stimuli and procedure were identical to those described in Study 2 with one exception. On the final

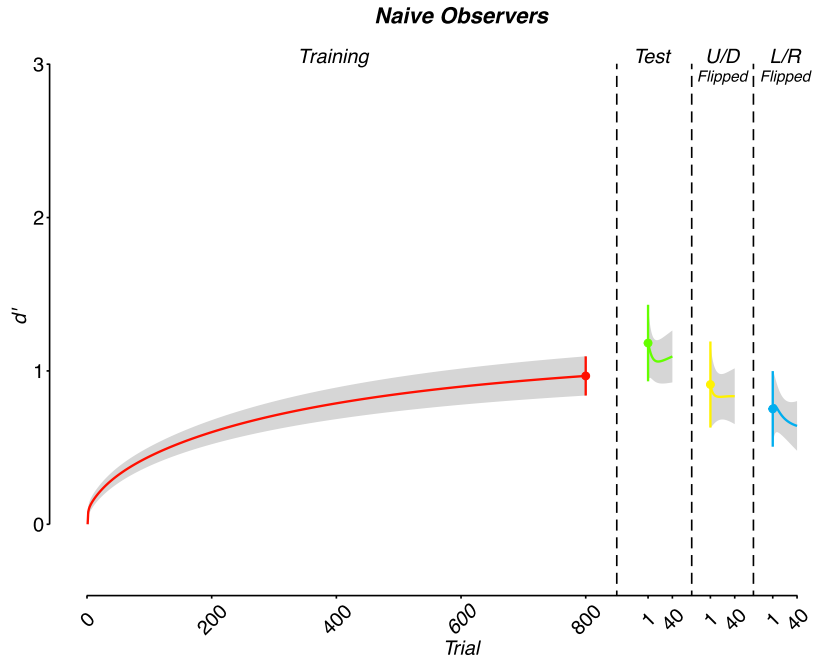

Figure 4. Naïve undergraduate performance. As was true of the trained psychophysical observers, the fully naïve undergraduate participants showed clear evidence of learning on the training task and full generalization to test images. In addition, reasonably full generalization was also seen to the flipped image versions.

day of the study, participants first completed the task on the test set of images (40 images alone, not intermixed with images from the training set as in Study 2). They then completed the same basic task on the flipped set of 80 images, which were the same flipped images used in Study 1; the up–down flipped images were presented first and then the left–right flipped images. Although the same informative feedback was provided during training as during Study 2, no feedback was provided on the final day where transfer was assessed (i.e., no feedback on test images or flipped images).

Results

Exclusions

A subset of 13 participants showed behavior indicating that they were not following task instructions (e.g., in one case the participant appeared to be immediately pressing the same button repeatedly as soon as each new trial appeared) or else did not complete all sessions. These participants were excluded from all further analyses; therefore, the results from a total of 49 participants were included.

Data analysis

Data analysis proceeded in the same manner as in Study 2.

Learning of trained images

As expected based on the results of Study 2, the participants showed clear evidence of learning, as their estimated final performance was significantly greater than 0: $M = 0.97 \pm 0.13$, $t(48) = 7.56$, $p < 0.001$, and $BF = 7.05$ (Figure 4).

Generalization to test images

Given that participants showed clear evidence of learning during training, we then examined the extent to which the learning generalized to the test images. As in Study 2, we first compared the participants' estimated starting performance on the test set of images during the final session to 0. The participants' estimated starting performance on the test images following training significantly exceeded 0: $M = 1.18 \pm 0.25$), $t(48) = 4.73$, $p < 0.001$, and $BF = 3.01$ (Figure 4). We then compared the participants' estimated starting performance on the test set of images to their estimated final performance on the training set of images. As expected, the participants' performance was not significantly different on the test and training sets of images: $t(48) = -0.92$, $p = 0.361$, and $BF = -0.63$ (Figure 4).

Generalization to flipped images

Finally, we conducted analogous analyses as those above, but with the two types of flipped images. A

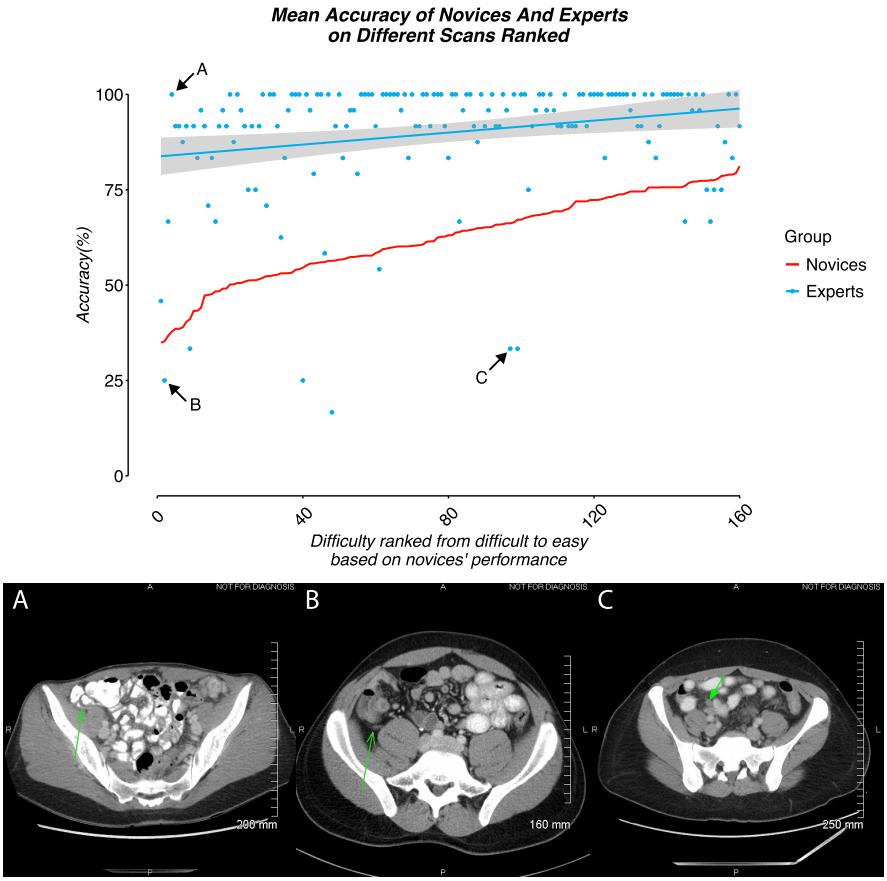

Figure 5. Comparison of expert and trained novice performance as a function of image. Images were ranked according to difficulty in the trained undergraduates. Although the experts substantially outperformed the trained novices (on almost every individual image), there is a clear correspondence between the images that experts and trained novices excelled at or struggled with. Below we provide examples of images (A) that experts were at ceiling performance on but which novices were very poor at; (B) that both experts and novices were poor at; and (C) that experts struggled disproportionately with as compared to novices.

significant difference in estimated starting performance was found comparing the up–down flipped images against 0: $M = 0.91 \pm 0.28$, $t(48) = 3.25$, $p = 0.002$, and $BF = 1.16$. No significant difference was found when comparing the up–down flipped image estimated starting performance with the estimated final performance of training: $t(48) = 0.21$, $p = 0.837$, and $BF = –0.80$ (Figure 4). Similarly, the estimated starting performance of left–right flipped images was significantly greater than 0: $M = 0.75 \pm 0.25$, $t(48) = 3.05$, $p = 0.004$, and $BF = 0.95$. It was not, however, significantly different from the estimated final performance of training sessions: $t(48) = 0.93$, $p = 0.358$, and $BF = –0.63$ (Figure 4).

Comparison to experts

In addition to our core questions regarding learning and generalization, an additional question of interest is whether participants appear to be learning a similar template as is utilized by expert performance (i.e., are they learning to evaluate images in a similar way). As one window into this question, we assessed whether there were similar patterns of errors in the experts and in the naïve observers post-training. If the groups were similar in terms of which images they found easy (and thus were highly accurate on) and which they found difficult (and thus were inaccurate on), it would suggest that similar knowledge was being applied to the problem. We thus ranked naïve participants' average performance (percent correct) on each image in the training set from difficult to easy (Figure 5, solid red line) and compared it to the average performance of experts on the same images (Figure 5, dots + blue line). We found that there was a significant correlation between the images that naïve participants and experts found easy or difficult with a Spearman's rank correlation test ($r_s = 0.22$, $p = 0.005$).

Discussion

The results of Study 3 replicated and extended those seen in Study 2. In Study 3, we confirmed that

the task was learnable even in a less selected group of participants and that the learning that emerged generalized to untrained images. In addition, there was clear generalization to up–down as well as left–right flipped images. This latter result is interesting in the context of the specificity of learning that has frequently been reported in the perceptual learning literature. We will return to this point in the Conclusions section.

Study 4: Is explicit feedback regarding the appendix location necessary for learning?

Given that the results of Study 3 showed that the task was learnable, generalized to new exemplars, and generalized to new orientations, in Studies 4 and 5 we sought to examine how aspects of the training task affect learning outcomes. Within the domain of perceptual learning, there is decidedly mixed evidence regarding whether explicit feedback is necessary to drive improvements in performance. In Studies 2 and 3, the feedback screen had two key components: (1) correct/incorrect feedback (i.e., whether or not the participant responded correctly), and (2) information regarding exactly where the appendix was in the image (the arrow pointing at the appendix). Although previous work examining the importance of feedback in driving perceptual learning has typically contrasted conditions with feedback against conditions without any feedback, here, given the highly complex nature of the stimuli, it was unclear whether the correct/incorrect feedback alone would be sufficient to drive learning over the time frame of the study. Indeed, in all models of learning, the purpose of feedback is to drive changes in estimates, models, templates, etc. In the case of simple tasks where participants are asked to make decisions about simple unitary dimensions, it is clear how correct/incorrect feedback alone could be sufficient (or in some cases not even necessary, such as if participants have an internal model that could produce an internal error signal) to drive learning. Yet, in the complex radiological stimulus space, we surmised that in the absence of understanding why choices were correct or incorrect, or where the relevant information was in the stimulus, learning may be significantly slowed or even eliminated, at least to the extent that we could measure this over five sessions of training. In Study 4, we thus sought to determine how necessary the second form of feedback that was provided in Studies 2 and 3, where participants were shown where the appendix had been in the previous image after each trial, is for the learning process.

Methods

Participants

Fifteen participants enrolled in an Introductory Psychology course participated in the study in exchange for course credit (mean age = 19.47 ± 2.42 years; 8 females). This sample size was consistent with a power analysis based on the results of Study 3; that is, given an expected effect size of $d' = 1.14$ and an alpha of 0.05, 14 participants (see below; one participant was excluded from the final analysis) would be associated with a power of greater than 0.95.

Stimuli and procedure

Participants underwent exactly the same procedure as in Study 3 with one exception. In the feedback screen that was seen after each trial during training, the arrow indicating the location of the appendix was omitted (the participants did receive the same explicit correct/incorrect feedback as in Studies 2 and 3).

Results

Exclusions

One participant was removed from analyses due to technical errors during data collection.

Data analysis

Data analysis proceeded in an identical fashion as in Study 3.

Learning of trained images

Although weak, participants did show some evidence of learning, as the estimated final performance exceeded 0: $M = 0.33 \pm 0.14$, $t(13) = 2.35$, $p = 0.036$, and $BF = 0.31$; (Figure 6).

Generalization to test images

The participants' estimated starting performance on the test images following training was not significantly different from 0: $M = 0.52 \pm 0.32$, $t(13) = 1.63$, $p = 0.126$, and $BF = –0.10$. It was also not significantly different from the estimated final performance on the training set: $t(13) = –0.67$, $p = 0.512$, and $BF = –0.48$.

Generalization to flipped images

Given that there was not significant evidence of any generalization to test images, we report the statistics for the flipped images simply for the sake of completeness. We did not observe a significant difference in the

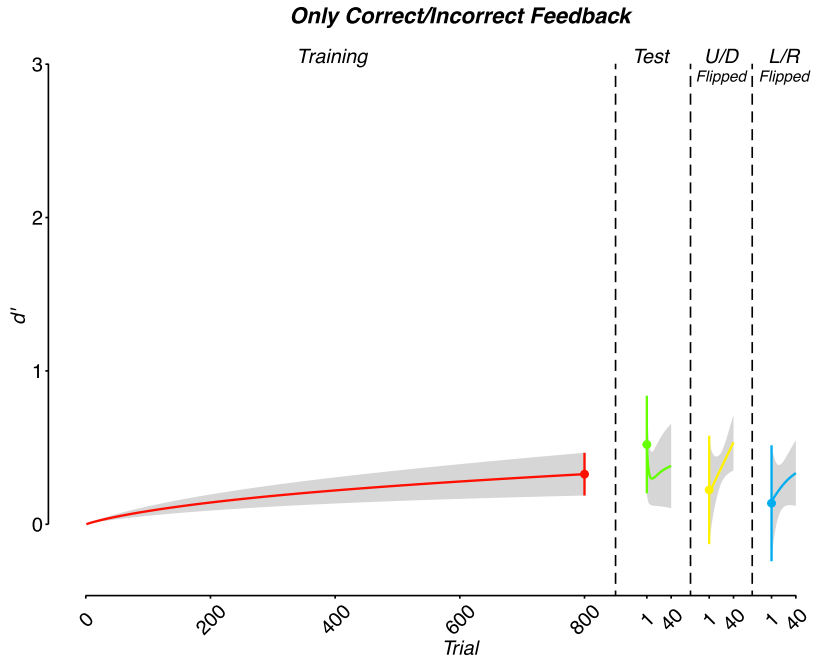

Figure 6. Performance when only correct/incorrect feedback was given during training. A weak, but statistically significant effect of training was found with participant performance at the conclusion of training exceeding chance levels. Yet, performance on the test/flipped images did not significantly exceed chance.

estimated starting performance of the up–down flipped images compared to either 0, where $M = 0.22 \pm 0.35$, $t(13) = 0.63$, $p = 0.538$, and $BF = –0.49$, or with the estimated final performance of the training session, where $t(13) = 0.28$, $p = 0.782$, and $BF = –0.55$. Similarly, the estimated starting performance on the left–right flipped images did not differ from 0: $M = 0.14 \pm 0.38$, $t(13) = 0.36$, $p = 0.723$, and $BF = –0.54$. Also, the estimated starting performance on the left–right flipped images did not differ significantly from the estimated final training session performance: $t(13) = 0.49$, $p = 0.631$, and $BF = –0.52$.

Comparison to Study 3

Qualitatively contrasting the results of Study 3 and Study 4 suggests the importance of explicit directed feedback in producing learning on the task; that is, learning and generalization to test images in Study 3 were significant when such feedback was provided, whereas in Study 4 we found significant learning but no significant generalization to the test images when such feedback was not provided. For this reason, we examined these findings more directly by contrasting the estimated final training performance between the two conditions. As expected, participants who received full feedback had an estimated final training d' significantly greater than that of those who received only correct/incorrect feedback: $t(37.03) = 3.39$, $p = 0.002$, and $BF = 1.40$. Note that here and in all between-groups comparisons, the degrees of freedom were calculated using the Satterhaite–Welch adjustment in order to be robust to unequal variances.

Discussion

The outcome of Study 4 demonstrates that, although binary feedback alone was perhaps sufficient to drive some learning (which was weak and perhaps specific to the trained image set), more detailed feedback, such as including information about where the appendix is located, is needed for learning to progress as was seen in Studies 2 and 3. The implications of this outcome will be discussed more fully in the Conclusions.

Study 5: Does ordering the training experience from easy to hard facilitate learning?

There is a great deal of evidence in various domains that the efficiency of the learning process is affected by the overall difficulty of the training task and how/whether this difficulty changes through time. Although the principle comes in various guises and with various names (e.g., scaffolding, zone of proximal development), in general, learning is often found to be most efficient when tasks become progressively more

difficult as participants learn and improve. Within the domain of perceptual learning, a huge proportion of perceptual learning studies are run utilizing staircases that naturally instantiate this principle. Indeed, when utilizing such staircases, participants are kept at a constant level of performance, such as 79% correct. To accomplish this, trials are continuously made more difficult in terms of the absolute stimulus space as participants become more adept at the task. Yet, in the case of more complex stimuli, it is unclear whether this simple approach—moving from easy to more difficult—will be as beneficial. In a standard perceptual learning task, such as orientation discrimination, there is a single dimension along which "easy" and "hard" trials fall. If the participants' goal is to differentiate between gratings oriented clockwise or counterclockwise from 45°, the same discriminant will be used for easy trials (where the gratings might differ by 15°) and for hard trials (where the gratings might differ by 3°). Yet, for a multidimensional stimulus such as was employed here, the space in which "easy" stimuli are found and the boundary that separates appendicitis from non-appendicitis within that space might be meaningfully different from that for "hard" stimuli. If this is true, then structuring training to move from easy to hard might have considerably less value. Here we sought to determine whether starting participants with easier examples and then later moving them to more difficult examples would benefit the learning process.

Methods

Participants

Fourteen participants enrolled in an Introductory Psychology course participated in the study in exchange for course credit (mean age = 19.07 ± 0.96 years; 8 females).

Stimuli and procedure

The methods and procedures were identical to those employed in Study 3 with the exception of the order in which the training images were presented. Based on the results of Study 3, images were ranked from easiest (i.e., highest overall percentage correct on the final day of training) to most difficult (i.e., lowest overall percentage correct on the final day of training). A median split was then utilized to create an "easy" training set and a "hard" training set. Participants encountered the easy set during the first two sessions of training. During the third session, half of the trials were from the easy set, and the remaining half were from the hard set. Finally, the hard set alone was used during the last two sessions of training.

Results

Exclusions

One participant was removed from analyses due to technical errors during data collection; one participant was removed for failing to complete all sessions. Therefore, a total of 12 participants were included in the final analyses.

Data analysis

We employed similar fitting methods as described under Studies 3 and 4. The one change was that the easy trials (first half of training) were fit separately from the hard trials (second half of training).

Learning of trained images

Participants showed clear evidence of learning on the easy trials; for estimated final performance, $M = 1.63 \pm 0.19$, $t(11) = 8.45$, $p < 0.001$, and $BF = 3.70$. Performance approached but did not reach significance for the hard trials: $M = 0.41 \pm 0.21$, $t(11) = 2.01$, $p = 0.070$, and $BF = 0.11$ (Figure 7).

Generalization to test images

The estimated starting performance of the test session was not significantly different from 0: $M = 0.95 \pm 0.64$, $t(11) = 1.5$, $p = 0.161$, and $BF = –0.15$.

Generalization to flipped images

The estimated starting performance of the up–down flipped images was not significantly different from 0: $M = 0.63 \pm 0.40$, $t(11) = 1.6$, $p = 0.138$, and $BF = –0.11$. The estimated starting performance of the left–right flipped images also did not differ significantly: $M = 1.09 \pm 0.52$, $t(11) = 2.1$, $p = 0.0595$, and $BF = 0.16$.

Comparison to Study 3

In order to compare performance in Study 3 with Study 5, we first separated training data in Study 3 based on whether the images were categorized as easy or hard in Study 5. Estimated final performance on the easy images did not differ between Studies 3 and 5: Study 3 $M = 1.72 \pm 0.22$ and Study 5 $M = 1.63 \pm 0.19$; $t(42.04) = 0.3$, $p = 0.767$, and $BF = –0.49$. Similarly, no difference was found for the hard images: Study 3 $M = 0.57 \pm 0.09$ and Study 5 $M = 0.41 \pm 0.21$; $t(15.07) = 0.68$, $p = 0.505$, and $BF = –0.42$. We also did not find a significant difference in estimated starting performance between the two studies in the test session: Study 3 $M = 1.18 \pm 0.25$ and Study 5 $M = 0.95 \pm 0.64$; $t(14.6) = 0.34$, $p = 0.741$, and $BF = –0.48$. However, we did find

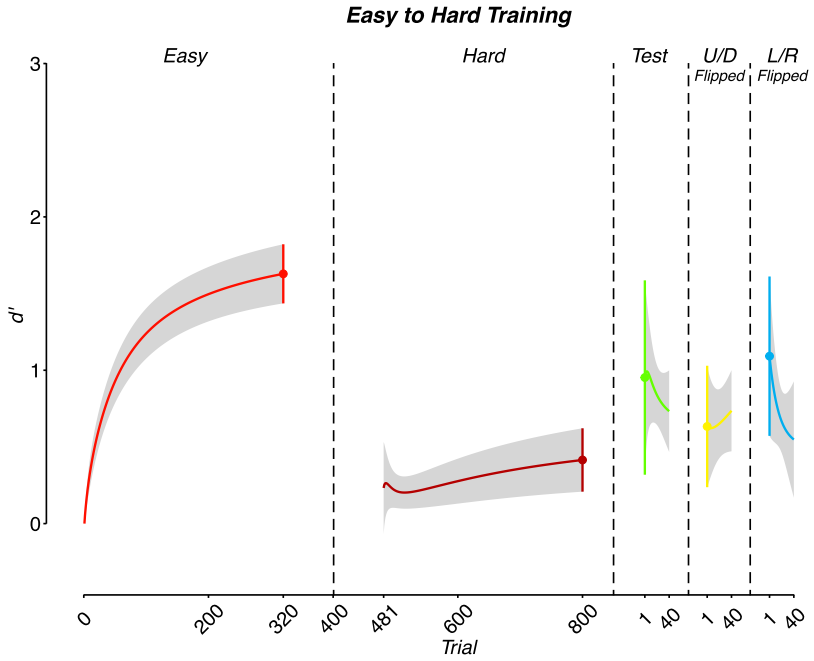

Figure 7. Performance when training was scaffolded from easy to hard images. As would be expected, participants learned quickly on the easy images during the early part of training; however, there was no reliable evidence of improvement on the hard images. Performance on the test and flipped images was reliably above chance (roughly in line with what was seen in Study 3).

that the estimated learning rate for Study 3 was greater than that for Study 5 for both easy and hard trials. For the easy trials, Study 3 $M = 6.93 \pm 0.44$ versus Study 5 $M = 5.32 \pm 0.59$; $t(25.12) = 2.19$, $p = 0.04$, and $BF = 0.30$. For the hard trials, Study 3 $M = 9.67 \pm 0.44$ versus Study 5 $M = 5.50 \pm 0.49$; $t(31.33) = 6.29$, $p < 0.001$, and $BF = 5.32$.

Discussion

In a manner counter to previous work showing that learning is more efficient when difficulty progresses from easy to harder trials, structuring training such that easy trials were experienced early in training and hard trials were experienced late in training appeared to have, at best, no significant impact on learning outcome. The only significant results were negative, in that the scaffolded training produced a reduced learning rate on test trials and performance on the test trials did not significantly exceed chance levels. We elaborate on these findings below.

Conclusions and future directions

Overall, the results show that the basic perceptual learning procedure employed in Studies 2 and 3 was successful in teaching novice participants to complete a difficult radiological classification task after only about 2 total hours of training. Importantly, the training gains appeared to generalize completely to novel images. This pattern of results rules out a learning strategy wherein participants simply memorized the answer associated with each unique image observed during training. Instead, it is consistent with the hypothesis that participants truly learned how to discriminate between the CT appearance of a normal appendix and appendicitis. This is further consistent with the work of Chen and colleagues who found little difference in learning to identify femur fractures as a function of whether or not cases were repeated during training (Chen et al., 2017). Beyond observing generalization to previously unseen test images, significant generalization was also observed to images presented in novel orientations. One possible explanation for the observed degree of learning generalization, which runs counter to much work in perceptual learning, is that this task involved a relatively high degree of variety in the stimuli (Schmidt & Bjork, 1992). The training set included 200 images, each with a unique anatomical location of the appendix and degree of associated inflammation (for those with appendicitis). By comparison, most cases of perceptual learning that have shown distinct task specificity tended to involve stimuli with considerably less variety (often only two exemplars or at a maximum exemplars differing along a single dimension).

Studies 4 and 5 addressed two key questions regarding training procedures. First, the results

of Study 4 demonstrated that, although corrective feedback alone was sufficient to drive some degree of learning, this learning was comparatively weak and perhaps specific to the trained images, as performance on the test images did not exceed chance level. Instead, it appears that what we dubbed "informative feedback"—feedback that indicated not only the correct answer but also the spatial location in the image where the most diagnostic information was present—is needed to see effective learning, at least over the time frame tested here. In addition to being important to the future development of perceptual training for radiological diagnoses, this observation provides interesting nuance to many of the debates regarding the importance and utility of feedback in perceptual learning. When utilizing the typical task stimuli and methods from much of the perceptual literature to date, participants should have minimal uncertainty regarding the spatial location of task-relevant information; that is, there is usually only a single stimulus and its position from trial to trial is known. Yet, in a more complex perceptual task such as the one used here, without knowing where the important information is would make it difficult to drive learning via only corrective feedback. In this study, much of the image was likely task irrelevant with regard to the appendicitis decision, but learning which part of the image was task relevant from only correct/incorrect feedback would be difficult. The results are therefore consistent with the general idea that an unnecessary situation is when the participants have a sufficiently well-developed internal model from which they can generate an internal error signal. In the absence of such a model (and internal error signal), informative feedback will be needed to drive learning (Dosher & Lu, 2017). This general view could also explain why Chen and colleagues saw significant learning in femur fracture identification with only corrective feedback (i.e., the equivalent of the feedback used in Study 4 here), as the femur itself is likely considerably easier to identify on an x-ray than the appendix on a single CT (Chen et al., 2017).

Finally, in contrast to the general trend in the learning literature, in Study 5 we found that structuring training experience from easy to difficult trials did not enhance learning. This, too, provides interesting nuance to discussions about properly structured or sequenced learning experiences. In a task where the stimuli vary along a single dimension, the same to-be-learned rule is important for both easy and hard trials; for example, if participants are attempting to indicate whether a Gabor is tilted clockwise or counterclockwise from 45°, learning what 45° looks like is important regardless of whether the stimuli are "easy" (as in ±15° from 45°) or "difficult"(as in ±1° from 45°). Yet, in a more complex stimulus space such as the one used here, it is entirely possible that the rule that best divides the space of stimuli into "yes" and "no" responses for the easy images differs in some fundamental ways from the rule that best divides the space for the hard images. If this is the case, by showing participants only easy images in the beginning, they may not be learning anything particularly useful regarding the way to approach the hard trials (in other words, the easy trials are not actually scaffolding the hard trials). The extent to which this is indeed the case is thus an interesting direction for future investigations (for evidence that over-representing difficult trials reduces learning, see Chen et al., 2017). However, it is critical to note that our method of titrating difficulty through time was quite coarse and not titrated to individual abilities. Although to some extent this was necessary given the current available data (e.g., our estimates of what trials are more or less difficult are similarly coarse), it leaves alive the possibility that more finely tuned changes in difficulty may produce enhancements in learning that were not observed with the methods employed here. We thus consider this to be a first step to examining the question, rather than a final answer. Gaining a deeper understanding of the dimensionality of the space and how the various dimensions impact difficulty (e.g., what exactly makes the hard images hard) may allow us to use more individualized and finely titrated methods in the future.

As noted earlier, these studies were meant as early examinations of the utility of using principles from the perceptual learning literature in training radiological diagnosis. Future investigations may find fruit in utilizing training that is more representative of actual clinical practice or examining the extent to which training generalizes to even more real-world contexts (van der Gijp, Ravesloot, Jarodzka, van der Schaaf, van der Schaaf, van Schaik, Ten Cate, 2017). For example, as noted earlier, clinical radiologists do not make diagnoses on the basis of a single image. Instead, they utilize the full complement of images through the abdomen, allowing them to manipulate all features including window, level, and image orientation (den Boer, van der Schaaf, Vincken, Mol, Stuijfzand, & van der Gijp, 2018; Drew, Vo, Olwal, Jacobson, Seltzer, & Wolfe, 2013). Thus, they need to be able to find the slices with the most diagnostic information for themselves; indeed, the search procedure itself appears to be a key part of radiological training. Whether training of the type provided here would result in enhancements when participants need to find relevant images, is thus unknown. Furthermore, in clinical practice, radiologists see mostly normal appendixes, instead of an equal distribution of normal and inflamed appendixes. By using an equal number of normal and abnormal cases during our training procedure, it is possible that we trained an incorrect bias. As such, testing participants with a more realistic distribution of positive and negative cases would be of interest. It is also the case that when making diagnoses radiologists do not only

rely on the visual images but also take into account non-imaging findings such as clinical presentation (e.g., patient is experiencing abdominal pain in a particular area), laboratory values, and previous history (e.g., recent vs. remote history of prior surgery). The impact of performing purely perceptual training in the absence of training in the use of other such sources of information is also important to determine. Finally, in real-world practice, radiologists have to evaluate for the full spectrum of diagnoses, not just appendicitis, which requires a much more comprehensive approach than simply identifying and categorizing the appendix. An ideal perceptual learning task would take each of these differences into account.

Keywords: perceptual learning, radiology, transfer of learning

Acknowledgments

Commercial relationships: none.
Corresponding author: C. Shawn Green.
Email: cshawn.green@wisc.edu.
Address: Department of Psychology, University of Wisconsin-Madison, Madison, WI, USA.

References

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

Bavelier, D., Bediou, B., & Green, C. S. (2018). Expertise and generalization: Lessons from action video games. *Current Opinion in Behavioral Sciences, 20*, 169–173.

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), e0189192.

Cochrane, A. (2020). *TEfits: Nonlinear regression for time-evolving indices*. Manuscript submitted for publication, https://doi.org/10.21105.joss.02535.

Cochrane, A., Cui, L., Hubbard, E. M., & Green, C. S. (2019). "Approximate number system" training: A perceptual learning approach. *Attention, Perception, & Psychophysics, 81*(3), 621–636.

den Boer, L., van der Schaaf, M. F., Vincken, K. L., Mol, C. P., Stuijfzand, B. G., & van der Gijp, A. (2018). Volumetric image interpretation in radiology: Scroll behavior and cognitive processes. *Advances in Health Sciences Education: Theory and Practice, 23*(4), 783–802.

De Valois, K. K. (1977). Independence of black and white: Phase-specific adaptation. *Vision Research, 17*(2), 209–215.

Deveau, J., & Seitz, A. R. (2014). Applying perceptual learning to achieve practical changes in vision. *Frontiers in Psychology, 5*, 1166.

Dosher, B., & Lu, Z. L. (2017). Visual perceptual learning and models. *Annual Review of Vision Science, 3*, 343–363.

Dosher, B. A., & Lu, Z. L. (2009). Hebbian reweighting on stable representations in perceptual learning. *Learning & Perception, 1*(1), 37–58.

Drew, T., Vo, M. L., Olwal, A., Jacobson, F., Seltzer, S. E., & Wolfe, J. M. (2013). Scanners and drillers: characterizing expert visual search through volumetric images. *Journal of Vision, 13*(10):3, 1–13, https://doi.org/10.1167/13.10.3.

Fahle, M. (1997). Specificity of learning curvature, orientation, and vernier discriminations. *Vision Research, 37*(14), 1885–1895.

Fahle, M., Edelman, S., & Poggio, T. (1995). Fast perceptual learning in hyperacuity. *Vision Research, 35*(21), 3003–3013.

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

Green, C. S., Banai, K., Lu, Z.-L., & Bavelier, D. (2018). Perceptual learning. In J. Serences (Ed.), *Steven's handbook of experimental psychology II: Sensation, perception & attention*. New York, NY: John Wiley & Sons.

Green, C. S., & Bavelier, D. (2008). Exercising your brain: A review of human brain plasticity and training-induced learning. *Psychology and Aging, 23*(4), 692–701.

Guadagnoli, M. A., & Lee, T. D. (2004). Challenge point: a framework for conceptualizing the effects of various practice conditions in motor learning. *Journal of Motor Behavior, 36*(2), 212–224.

Herzog, M. H., & Fahle, M. (1997). The role of feedback in learning a vernier discrimination task. *Vision Research, 37*(15), 2133–2141.

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

Kattner, F., Cochrane, A., Cox, C. R., Gorman, T. E., & Green, C. S. (2017). Perceptual learning generalization from sequential perceptual training as a change in learning rate. *Current Biology, 27*(6), 840–846.

Kattner, F., Cochrane, A., & Green, C. S. (2017). Trial-dependent psychometric functions accounting for perceptual learning in 2-AFC discrimination tasks. *Journal of Vision, 17*(11):3, 1–16, https://doi.org/10.1167/17.11.3.

Kellman, P. J. (2013). Adaptive and perceptual learning technologies in medical education and training. *Military Medicine, 178*(10, suppl.), 98–106.

Kelly, B., Rainford, L. A., McEntee, M. F., & Kavanagh, E. C. (2018). Influence of radiology expertise on the perception of nonmedical images. *Journal of Medical Imaging, 5*(3), 031402.

Kelly, B. S., Rainford, L. A., Darcy, S. P., Kavanagh, E. C., & Toomey, R. J. (2016). The development of expertise in radiology: In chest radiograph interpretation, "expert" search pattern may predate "expert" levels of diagnostic accuracy for pneumothorax identification. *Radiology, 280*(1), 252–260.

Kundel, H. L., & Nodine, C. F. (1983). A visual concept shapes image perception. *Radiology, 146*(2), 363–368.

Li, Z. S., 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.

Liu, J., Lu, Z. L., & Dosher, B. A. (2010). Augmented Hebbian reweighting: Interactions between feedback and training accuracy in perceptual learning. *Journal of Vision, 10*(10):29, 1–14, https://doi.org/10.1167/10.10.29.

McKee, S. P., & Westheimer, G. (1978). Improvement in vernier acuity with practice. *Perception & Psychophysics, 24*(3), 258–262.

Petrov, A. A., Dosher, B. A., & Lu, Z. L. (2006). Perceptual learning without feedback in non-stationary contexts: Data and model. *Vision Research, 46*(19), 3177–3197.

Poggio, T., Fahle, M., & Edelman, S. (1992). Fast perceptual learning in visual hyperacuity. *Science, 256*(5059), 1018–1021.

Ravesloot, C. J., van der Schaaf, M. F., Kruitwagen, C., van der Gijp, A., Rutgers, D. R., Haaring, C., … van Schaik, J. P. J. (2017). Predictors of knowledge and image interpretation skill development in radiology residents. *Radiology, 284*(3), 758–765.

Saffell, T., & Matthews, N. (2003). Task-specific perceptual learning on speed and direction discrimination. *Vision Research, 43*(12), 1365–1374.

Sagi, D. (2011). Perceptual learning in vision research. *Vision Research, 51*(13), 1552–1566.

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science, 3*(4), 207–217.

Seitz, A. R. (2017). Perceptual learning. *Current Biology, 27*(13), R631–R636.

Seitz, A. R., Nanez Sr., J. E., Holloway, S., Tsushima, Y., & Watanabe, T. (2006). Two cases requiring external reinforcement in perceptual learning. *Journal of Vision, 6*(9), 966–973, https://doi.org/10.1167/6.9.9.

Sowden, P. T., Davies, I. R., & 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.

Sowden, P. T., Rose, D., & Davies, I. R. (2002). Perceptual learning of luminance contrast detection: Specific for spatial frequency and retinal location but not orientation. *Vision Research, 42*(10), 1249–1258.

van der Gijp, A., Ravesloot, C. J., Jarodzka, H., van der Schaaf, M. F., van der Schaaf, I. C., van Schaik, J. P. J., … Ten Cate, Th. J. (2017). How visual search relates to visual diagnostic performance: A narrative systematic review of eye-tracking research in radiology. *Advances in Health Sciences Education: Theory and Practice, 22*(3), 765–787.

Vogels, R., & Orban, G. A. (1986). Decision processes in visual discrimination of line orientation. *Journal of Experimental Psychology: Human Perception and Performance, 12*, 115–132.

Vygotsky, L. S. (1978). *Mind and society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Waite, S., Farooq, Z., Grigorian, A., Sistrom, C., Kolla, S., Mancuso, A., … Macknik, S. L. (2020). A review of perceptual expertise in radiology: How it develops, how we can test it, and why humans still matter in the era of artificial intelligence. *Academic Radiology, 27*(1), 26–38.

Zhang, E., & Li, W. (2010). Perceptual learning beyond retinotopic reference frame. *Proceedings of the National Academy of Sciences, USA, 107*(36), 15969–15974.