Journal of Applied Research in Memory and Cognition xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Journal of Applied Research in Memory and Cognition

journal homepage: www.elsevier.com/locate/jarmac

Improved classification of mammograms following idealized training

Adam N. Hornsby, Bradley C. Love*

Experimental Psychology, University College London, 26 Bedford Way, London WC1H 0AP, United Kingdom

ARTICLE INFO

Article history:
Received 2 February 2014
Received in revised form 22 April 2014
Accepted 24 April 2014
Available online xxx

Keywords:
Categorization
Memory retrieval
Decision making
Mammograms
Idealization
Medical diagnosis

ABSTRACT

People often make decisions by stochastically retrieving a small set of relevant memories. This limited retrieval implies that human performance can be improved by training on idealized category distributions (Giguère & Love, 2013). Here, we evaluate whether the benefits of idealized training extend to categorization of real-world stimuli, namely classifying mammograms as normal or tumorous. Participants in the idealized condition were trained exclusively on items that, according to a norming study, were relatively unambiguous. Participants in the actual condition were trained on a representative range of items. Despite being exclusively trained on easy items, idealized-condition participants were more accurate than those in the actual condition when tested on a range of item types. However, idealized participants experienced difficulties when test items were very dissimilar from training cases. The benefits of idealization, attributable to reducing noise arising from cognitive limitations in memory retrieval, suggest ways to improve real-world decision making.

1. Introduction

Classifying mammograms as tumorous vs. non-tumorous is a complex, probabilistic, and error-prone task. When performing such tasks, one possibility is that people selectively and stochastically retrieve relevant memories to guide their decisions (Giguère & Love, 2013; Nosofsky & Palmeri, 1997). Unfortunately, selectively sampling memory introduces noise at the time of decision that results in suboptimal performance (Giguère & Love, 2013).

Recently, there has been interest in tailoring training conditions to promote better test performance (Pashler & Mozer, 2013). For instance, Giguère and Love (2013) find that the harmful effects caused by limited memory retrieval at the time of decision can be reduced by training people on idealized distributions of category information. This idealization, which deemphasizes ambiguous cases, reduces the likelihood that misleading memories will be retrieved at the time of test. For example, in one study, two groups were trained on a random set of baseball games and asked to predict at test the outcomes of the remaining games for that season. The group trained on the actual outcomes of the games did not perform as well at test as the group trained on idealized game outcomes based on the total wins by teams in the training set.

Idealization may be complementary to other techniques that aim to improve human performance given limits in cognitive abilities. For example, manipulating the presentation order of training examples is another method to make the underlying structure of information more salient. Presentation orders that make the category structure more salient lead to better learning (Avrahami et al., 1997; Clapper & Bower, 1994; McClelland, Fiez, & McCandliss, 2002; Medin & Bettger, 1994). For example, people are better able to separate contrasting structures when a number of items from one category are presented together followed by a number of contrasting items from the other category (Clapper & Bower, 1994). These ordering effects, which isolate the categories (cf. Goldstone, 1996), make the contrasting structure evident. Other order manipulations emphasize presenting unambiguous cases (from either category) early in learning and only presenting the ambiguous cases later in learning after the learner is properly anchored (Avrahami et al., 1997; McClelland et al., 2002). These ordering manipulations that promote structure discovery share a kindred spirit with idealization. Whereas an advantageous item ordering makes the underlying category structure more apparent by strengthening contrast, idealization of category structures increases contrast by removing or altering ambiguous cases.

One key question is whether the idealization advantage extends to classification tasks that involve complex real-world stimuli. Extending to complex real-world stimuli would bring the idealization manipulation one step closer to useful application. Previous results in the literature suggest that idealization and the psychological theory underlying it should extend to real-world settings. Even

* Corresponding author. Tel.: +44 0207 679 1515.
E-mail addresses: adam.hornsby.10@alumni.ucl.ac.uk (A.N. Hornsby), b.love@ucl.ac.uk (B.C. Love).

http://dx.doi.org/10.1016/j.jarmac.2014.04.009

when provided with explicit instruction, dermatological diagnoses are guided by similarity to experienced examples (Chan, Brooks, & Norman, 2001). In a simulated psychiatric diagnosis task, participants relied on easily accessible instances and their decisions were guided by the idiosyncratic properties of the training items (Young, Brooks, & Norman, 2011). These results align closely with Giguère and Love (2013) characterization of memory retrieval at the time of decision. To the extent that memory retrieval is limited to available (i.e., recent, familiar, and similar) instances, idealized training should improve test performance.

Here, we examine whether idealized training improves people's ability to classify novel mammograms as tumorous or non-tumorous (i.e. normal). In Experiment 1, we normed mammograms to determine the a priori ambiguity or difficulty level (easy, medium, or hard) of the images. In the main study, Experiment 2, a second group of participants were trained to classify mammograms using trial-and-error learning (i.e. stimulus → response → feedback). We correctly predicted that participants trained on an idealized distribution of mammograms (i.e. only including unambiguous easy cases) would be more accurate in classifying novel mammograms (across difficulty levels) at test than participants trained on a representative distribution of mammograms that included easy, medium, and hard items as in the test set. However, the results were nuanced in that the idealization advantage was strongest for images that were somewhat similar to those experienced during training.

2. Experiment 1

Unlike the simple stimuli typically used in category learning studies, mammograms are subtle, complex, and high-dimensional. These stimuli are not easily described in terms of basic stimulus dimensions (e.g., size, shape, and color) that are psychologically meaningful to participants. Rather than attempt to discover the dimensional structure of mammograms, which may be an intractable task and is likely not agreed upon across individuals, the goal of Experiment 1 is to norm mammograms to determine prior to training how likely people are to view an image as containing a tumor. These stimulus ratings are used in the main study, Experiment 2.

2.1. Method

2.1.1. Participants

One hundred participants were recruited using Amazon Mechanical Turk (mturk). Mturk (www.mturk.com) has been used for a wide variety of psychological studies and has been shown to be an inexpensive, fast, and reliable source of human data (Buhrmester, Kwang, & Gosling, 2011; Crump, McDonnell, & Gureckis, 2013). All participants were required to have had 90% or more of their previous mturk assignments approved. Eighteen participants were removed from the final analyses because they failed two or more of the catch trials (described below). The mean age of the final sample was 33.0 (SD = 28.9). Participants were from 12 different countries with 85.37% either from the USA or India. Participants were paid $.50 for participating, and were awarded an additional $.50 bonus for correctly responding in all five catch trials. This pay level is typical for mturk (Horton & Chilton, 2010).

2.1.2. Apparatus and stimuli

The study was designed using Adobe's ActionScript language and was accessed using Adobe Flash Player in a web browser. The task was presented in a black window, which was 600 × 450 pixels. Participants responded by clicking on a green 'Normal' button on the left of the window or a red 'Tumour' button on the right with their computer mouse. All mammograms were at a mediolateral oblique (MLO) angle, left-facing, taken with a 43.5 μm HOWTEK scanner, and presented in the lossless Portable Network Graphics (PNG) format. Example images are shown in Fig. 1.

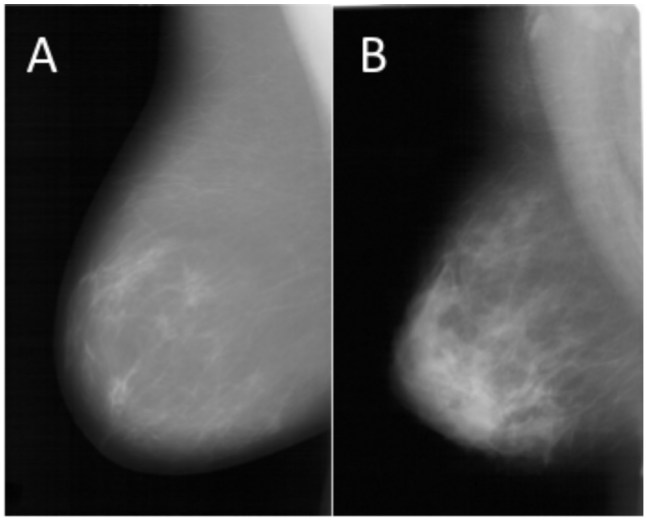

Fig. 1. A normal (A) and tumorous (B) mammogram.

2.1.3. Procedure

Participants were asked to enter their age, sex and location, and to confirm that they had no prior medical training or professional experience with classifying mammograms. On each of the 200 rating trials, 3 images were randomly selected from the bank of 358 possible images, subject to the constraint that the left image was normal, the right image was tumorous, and all 3 images were unique. The participant's task was to decide whether the central image was normal or contained a tumor. The three images were presented for 2000 ms before the response buttons appeared. Images were presented until a response was made and then a blank screen was shown for 1000 ms (no corrective feedback was provided). In addition to the 200 rating trials, there were five randomly interspersed catch-trials in which the image in the center was identical to one of the flanking images, which should make the correct response clear. A progress bar was displayed at the top of the screen and participants were fully debriefed at the end the study.

2.2. Results and discussion

Fig. 2 summarizes the rating data. The results confirm that images vary greatly in their a priori difficulty with some images being very misleading and hard to classify. The percentage of correct responses made to each image in the norming study (see Fig. 2)

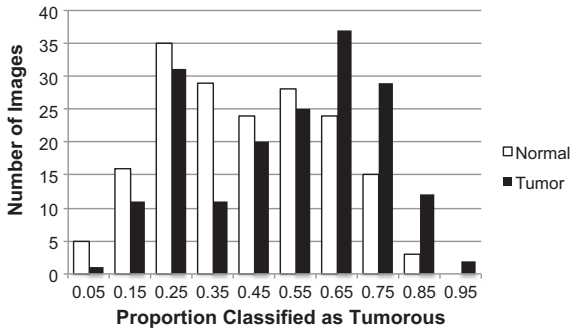

Fig. 2. Participants' distribution of tumor judgments in the norming task for normal (white) and tumorous (black) mammograms. Each bin is centered on the value shown beneath it.

Please cite this article in press as: Hornsby, A. N., & Love, B.C. Improved classification of mammograms following idealized training. *Journal of Applied Research in Memory and Cognition* (2014), http://dx.doi.org/10.1016/j.jarmac.2014.04.009

Table 1
The proportion of normal and tumorous images classified as easy, medium and hard as a result of participants' responses in the norming study.

Image type	Range of correct responses (%)	Normal images (%)	Tumor images (%)
Easy	67–100	37	35
Medium	51–66	23	23
Hard	0–50	39	42

was used to indicate how difficult (easy, medium, or hard) they were to classify prior to training. The three splits were greater than 66%, 50–66%, and less than 50% for easy, medium, and hard, respectively. The proportion of images placed in to each grouping is shown in Table 1.

3. Experiment 2

Working with the normed mammograms from Experiment 1, participants trained on either a representative set of easy, medium, and hard images (see Table 1) in the actual condition or on a set of exclusively easy images in the idealized conditions. Both groups were tested on a novel set of representative stimuli (i.e., easy, medium, and hard images). As discussed in Section 1, idealized training is predicted to have benefits because it reduces noise introduced by limits in memory retrieval at the time of decision. If so, participants in the idealized condition should outperform those in the actual condition.

3.1. Method

3.1.1. Participants

Participants were recruited online using Amazon Mechanical Turk. This study demanded sustained concentration from participants. Participants failing to meet certain compliance conditions were excluded from further analyses. Specifically, participants were removed if they exited full-screen mode, had a training- or test-phase response-time mean more than two standard deviations from the mean across all participants, or repeated or alternated responses ten times. Using these criteria, 54 of the 265 participants were removed. The final sample consisted of 211 participants (110 = idealized, 101 = actual) with 94% from the USA or India. Participants were paid $1.00 for participating and earned a bonus of $.50 if their test phase accuracy was above 60%. The participant with the highest accuracy was awarded an additional $10 bonus.

3.1.2. Procedure and stimuli

The study was accessed using Adobe Flash Player in a web browser and required full screen throughout. *Normal* and *Tumor* responses were made with the left and right arrow keys of the keyboard, respectively.

The stimuli comprised 358 normal and tumorous mammograms taken from the Digital Database for Screening Mammography (DDSM; Heath, Bowyer, Kopans, Moore, & Kegelmeyer, 2001). The image's respective width and height was always 38.5% × 64.8% of the participant's screen. The training set consisted of 108 unique stimuli selected randomly (all randomization done per participant) with constraints (described below) from the 358 possible stimuli. The training phase consisted of three trial blocks in which each of the 108 stimuli were presented once in a random order.

Participants were randomly assigned to either the idealized or actual training condition. In the idealized condition, the 108 stimuli consisted of 54 easy images for each category (normal or tumorous). In the actual condition, the 108 stimuli consisted of 18 easy, medium, and hard images for each category. On each training trial, the stimulus was shown, participants responded, corrective feedback (with image still shown) indicating tumorous or normal was provided for 2000 ms, and then a blank screen was shown for 500 ms.

Table 2
A signal detection analysis of participants' responses during the test phase for each condition and image type.

Condition	Image type	Hits	False alarms	d'	Criterion
Actual	Easy	0.81	0.26	1.523	−0.103
Idealized	Easy	0.97	0.09	3.183	−0.256
Actual	Medium	0.65	0.42	0.582	−0.098
Idealized	Medium	0.78	0.44	0.922	−0.303
Actual	Hard	0.38	0.68	−0.774	−0.071
Idealized	Hard	0.19	0.88	−2.083	−0.148

After completing all 324 training trials, participants completed 18 test trials, which consisted of three previously unseen easy, medium and hard items from each category displayed in a random order. Test trials followed the same procedure as training trials, except 'Thank You' was displayed instead of corrective feedback.

3.2. Results and discussion

Before presenting the main analyses of test performance, we begin with an analysis of performance on training trials. A 3 × 2 mixed factorial ANOVA assessed accuracy over the three training blocks in the actual and idealized conditions. Participants improved over training blocks, $F(2,415) = 15.10$, $p < .01$, $\eta_p^2 = .066$,[1] with a positive linear trend, $F(1,214) = 24.82$, $p < .01$, $\eta_p^2 = .104$. Overall, participants were more accurate (.92 vs. .58) in the idealized condition, $F(1,214) = 9720.99$, $p < .01$, $\eta_p^2 = .978$. Block and training condition interacted such that participants in the idealized condition showed greater improvement over blocks, $F(2,415) = 11.04$, $p < .01$, $\eta_p^2 = .049$. Indeed, participants in the actual condition did not show strong gains across blocks – a post hoc analysis of the three items types showed no differences in performance between the first and last block for the easy and medium items, $t(100) = 1.10$, $p = .273, = .11$ and $t(100) = .67$, $p = .507, = .007$, but did find an increase in performance (.33 vs. .36) for hard items, $t(100) = 2.87$, $p < .01$.

The main result (see Fig. 3A) was that participants were more accurate in classifying novel test items in the idealized (.60, $SE = 84$) than the actual (.57, $SE = 1.0$) condition, $t(209) = 2.14$, $p = .034$, $d = 0.30$. Analyzing by image type (see Fig. 3B), the idealized condition was advantaged for easy and medium items, $t(155) = 9.82$, $p < .001$, $d = 1.42$; $t(194) = 2.92$, $p = .004$, $d = 0.41$, whereas the actual was advantaged for hard items, $t(177) = 7.64$, $p < .001$, $d = 1.07$.[2] Notice that participants in the actual condition are closer to chance guessing in all conditions. This result is anticipated by the theory and analyses presented by Giguère and Love (2013), which conclude that samples from memory should be less informative and noisy in the actual than the idealized condition.

These test results invite additional scrutiny, particularly the reversal of the idealization advantage for hard items. This difference in accuracy for hard items across conditions is not readily attributed to a difference in response bias as the proportion of tumor judgments was similar in the actual (.54) and idealized (.52) conditions, $F(1,209) = .748$, $p = .388, = .004$. Indeed, a signal detection analysis (see Table 2) that aggregated over participants'

[1] Greenhouse-Geisser's probability values and degrees of freedom are reported for all within-subjects main effects and interactions in the training-phase data, as the Mauchly's test for violation of the sphericity assumption was significant ($p = .035$).

[2] Degrees of freedom in each comparison were corrected for inequality of variances. Levene's test results: Easy images: F = 32.24, p < .001, Medium: F = 5.71, p = .018, Hard: F = 14.46, p < .001.

Please cite this article in press as: Hornsby, A. N., & Love, B.C. Improved classification of mammograms following idealized training. *Journal of Applied Research in Memory and Cognition* (2014), http://dx.doi.org/10.1016/j.jarmac.2014.04.009

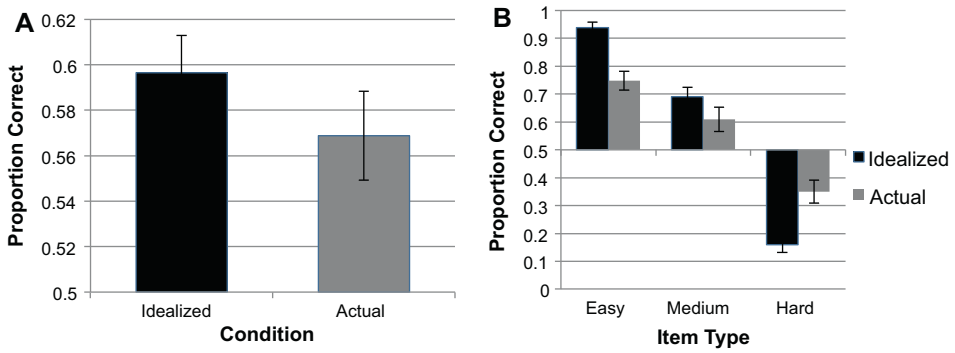

Fig. 3. Test accuracy. Participants were more accurate overall at test in the idealized condition (Panel A). This advantage held for easy and medium items, but not for hard items where both conditions were below chance (Panel B). Error bars are 95% confidence intervals.

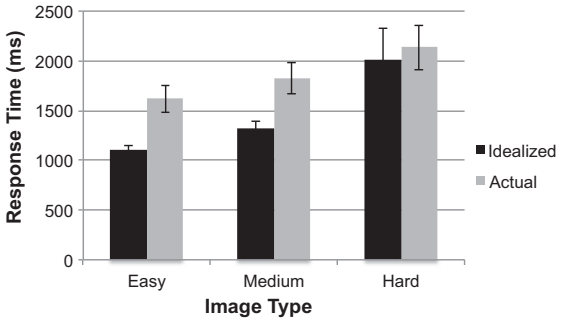

Fig. 4. Test response time. Participants in the actual condition made slower responses when correct for easy and medium images, but not for hard images, where there was no significant difference between the idealized and actual condition. Error bars are 95% confidence intervals.

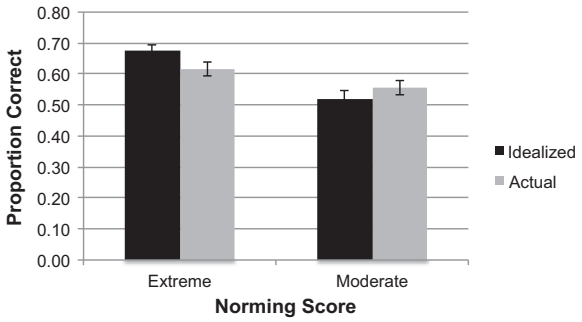

Fig. 5. Test accuracy for extreme and moderate items. Participants in the idealized condition made more accurate responses for extreme items (normed as less than .33 or greater than .66), whereas those in the actual condition made more accurate responses for items within that range. Error bars are 95% confidence intervals.

responses found that participants in both conditions had a slight tendency to respond "tumor" (i.e., their decision criterion was liberal). Response bias alone does not explain the pattern of results.

Although confidence ratings were not collected, response times at test provide may provide a surrogate measure of confidence. Response times for correct responses are typically slower when participants have low confidence (Yeung & Summerfield, 2012). A 2×3 mixed factorial ANOVA was conducted to assess participants' median response times for correct responses (see Fig. 4) for the three item types. There was a significant main effect of image type, $F(1,255) = 45.64$, $p < .001$, $\eta_p^2 = .190$.[3] Analyses also revealed that median response times for correct responses were significantly slower in the actual condition (1861 ms vs. 1470 ms) than the idealized condition, $F(1,195) = 15.22$, $p < .001$, $\eta_p^2 = .072$. There was also a significant interaction between difficulty and condition, $F(1,255) = 4.49$, $p = .025$, $\eta_p^2 = .022$. The interaction indicates that participants in the idealized condition slowed down more for hard items, which is consistent with the idea that participants may be aware of their uncertainty. Overall, response times roughly mirrored the accuracy data.

The previous analyses indicate an overall advantage for idealized training, but with deleterious effects for hard items. The results are not readily explained by a biasing account. One possible explanation for the pattern of results is that idealized training for certain image classes does not readily extend to all classes. Participants in the idealized condition only trained on the easy item type. The images for easy items were rated (see Fig. 2 and Table 1) in Experiment 1 as being clearly normal (i.e., the proportion of participants rating these images as tumorous was .33 or less) or clearly tumorous (i.e., the proportion of participants rating these images as tumorous was .66 or more). In other words, participants in the idealized condition only trained on extreme images, whereas those in the actual condition experienced the entire range.

One possibility is that extreme images have characteristics in common that do not apply to more moderate images. If so, one would expect that there would be an interaction such that participants in the idealized condition should outperform those in the actual condition on extreme items (normed as less than .33 or greater than .66) whereas participants in the actual condition should perform better on the images with moderate ratings (normed from .33 to .66). This pattern of results held (see Fig. 5) and the interaction was significant, $F(1,209) = 14.35$, $p < .01$, $\eta_p^2 = .064$.

4. General discussion

Idealized training does improve people's ability to categorize real-world stimuli such as mammograms. Strikingly, participants in the idealized condition were more accurate at classifying medium-difficulty images even though they were never exposed to this class of stimuli in training. However, idealized participants performed worse on hard images, though both groups were below chance.

These hard images may be truly misleading. On occasions, even medical doctors rely on additional tests to determine the presence and nature of a tumor. Indeed, the correct classifications of mammograms in the present study were confirmed using physical examinations or biopsy. In real-world practice, 57% of breast cancers are identified through mammography alone with

[3] Greenhouse-Geisser's probability values and degrees of freedom are reported for all within-subjects main effects and interactions in the training-phase data, as the Mauchly's test for violation of the sphericity assumption was significant ($p < .001$).

Please cite this article in press as: Hornsby, A. N., & Love, B.C. Improved classification of mammograms following idealized training. *Journal of Applied Research in Memory and Cognition* (2014), http://dx.doi.org/10.1016/j.jarmac.2014.04.009

the remaining 43% requiring further screening to make a diagnosis (Mathis et al., 2010). One possibility is that the hard items would fall within this 43% demanding further scrutiny. For instance, if idealized learners were given the option of flagging problematic items, perhaps they would have indicated that they were uncertain in their diagnosis of these hard items. Future studies will elicit confidence ratings to directly evaluate this possibility, but the present study offers some support for this possibility – response time patterns indicate that idealized learners were uncertain about their decisions for hard items.

One key for future research is to understand how such manipulations affect expert participants (e.g., radiologists) and the development of expertise. Although experts should face the same memory retrieval limitations as novices, experts represent mammograms in a richer fashion. Idealized training or refreshers for experts should take into account their richer knowledge. Related, the development of expertise may require exposure to a range of stimuli. This need must be balanced with the benefits of idealization, which restricts exposure to ambiguous cases. The present results are supportive of this conclusion. Participants in the idealized condition were advantaged only on stimuli that were somewhat similar (in terms of the norming score) to stimuli experienced during training.

Likewise, training procedures need to take into account the costs of different types of errors. For example, the cost of classifying a tumorous mammogram as normal is likely greater than the converse. In such cases, training should bias away from high-cost errors. Signal detection analyses in the present study did not reveal any systematic group differences in response bias for actual and idealized participants.

4.1. Practical application

To summarize the practical lessons from these studies, learners (e.g., radiologists) should be trained on idealized distributions of information which minimize the saliency of ambiguous cases, but care should also be taken to expose the learner to an adequate range of stimuli. The present study likely overly restricted the range of experiences. Rather than train only on easy items, one recommendation is to train on easy and medium items (only omitting hard items), which would cover the entire range of stimuli when both the normal and tumorous categories are taking into account. This approach may harness the benefits of idealization while minimizing the drawbacks. Finally, measures of confidence, such as response time and explicit ratings, should be used to determine which items merit additional scrutiny.

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgements

We are grateful to T.M. Deserno, Department of Medical Informatics, RWTH Aachen, Germany for providing PNG format images of the DDSM database.

References

Avrahami, J., Kareev, Y., Bogot, Y., Caspi, R., Dunaevsky, S., & Lerner, S. (1997). Teaching by examples: Implications for the process of category acquisition. *Quarterly Journal of Experimental Psychology*, *50*, 586–606.

Buhrmester, M., Kwang, T., & Gosling, S. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3–5. http://dx.doi.org/10.1177/1745691610393980

Chan, K-M., Brooks, L. R., & Norman, G. R. (2001). Coordination of analytic and similarity-based processing strategies and expertise in dermatological diagnosis. *Teaching and Learning in Medicine*, *13*(2), 110–116.

Clapper, J. P., & Bower, G. H. (1994). Category invention in unsupervised learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *20*, 443–460.

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS ONE*, *8*(3), e57410. http://dx.doi.org/10.1371/journal.pone.0057410

Giguère, G., & Love, B. (2013). Limits in decision making arise from limits in memory retrieval. *Proceedings of the National Academy of Sciences*, *110*(19), 7613–7618. http://dx.doi.org/10.1073/pnas.1219674110

Goldstone, R. L. (1996). Isolated and interrelated concepts. *Memory & Cognition*, *24*, 608–628.

Heath, M., Bowyer, K., Kopans, D., Moore, R., & Kegelmeyer, W. (2001). The digital database for screening mammography. In *Proceedings of the fifth international workshop on digital mammography* (pp. 212–218). Medical Physics Publishing.

Horton, J., & Chilton, L. (2010). The labor economics of paid crowdsourcing. In *Proceedings of the 11th ACM conference on electronic commerce. SSRN*.

Mathis, K. L., Hoskin, T. L., Boughey, J. C., Crownhart, B. S., Brandt, K. R., Vachon, C. M., et al. (2010). Palpable presentation of breast cancer persists in the era of screening mammography. *Journal of the American College of Surgeons*, *210*(3), 314–318.

McClelland, J. L., Fiez, J. A., & McCandliss, B. D. (2002). Teaching the/r/-/l/discrimination to Japanese adults: Behavioral and neural aspects. *Physiology and Behavior*, *77*(4–5), 657–662.

Medin, D. L., & Bettger, J. G. (1994). Presentation order and recognition of categorically related examples. *Psychonomic Bulletin & Review*, *1*, 250–254.

Nosofsky, R., & Palmeri, T. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review*, *104*, 266–300.

Pashler, H., & Mozer, M. C. (2013). When does fading enhance perceptual category learning? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(4), 1162–1173.

Yeung, N., & Summerfield, C. (2012). Metacognition in human decision making: Confidence and error monitoring. *Philosophical Transactions of the Royal Society B*, *367*, 1310–1321.

Young, M. E., Brooks, L. R., & Norman, G. R. (2011). The influence of familiar non-diagnostic information on the diagnostic decisions of novices. *Medical Education*, *45*(4), 407–414.

Please cite this article in press as: Hornsby, A. N., & Love, B.C. Improved classification of mammograms following idealized training. *Journal of Applied Research in Memory and Cognition* (2014), http://dx.doi.org/10.1016/j.jarmac.2014.04.009