Connecting Adaptive Perceptual Learning and Signal Detection Theory in Skin Cancer Screening

Philip J. Kellman (kellman@cognet.ucla.edu)[1,3]
Christine M. Massey (cmassey@psych.ucla.edu)[1]
Sally Krasne (skrasne@mednet.ucla.edu)[2]
Everett W. Mettler (mettler@ucla.edu)[1]

[1]Department of Psychology, University of California, Los Angeles, Los Angeles, CA 90095, USA
[2]Department of Physiology and [3]Department of Surgery,
David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA

Abstract

Combining perceptual learning techniques with adaptive learning algorithms has been shown to accelerate the development of expertise in medical and STEM learning domains (Kellman & Massey, 2013; Kellman, Jacoby, Massey & Krasne, 2022). Virtually all adaptive learning systems have relied on simple accuracy data that does not take into account response bias, a problem that may be especially consequential in multi-category perceptual classifications. We investigated whether adaptive perceptual learning in skin cancer screening can be enhanced by incorporating signal detection theory (SDT) methods that separate sensitivity from criterion. SDT-style concepts were used to alter sequencing, and separately to define mastery (category retirement). SDT retirement used a running d' estimate calculated from a recent window of trials based on hit and false alarm rates. Undergraduate participants used a Skin Cancer PALM (perceptual adaptive learning module) to learn classification of 10 cancerous and readily-confused non-cancerous skin lesion types. Four adaptive conditions varied either the type of adaptive sequencing (standard vs. SDT) or retirement criteria (standard vs. SDT). A non-adaptive control condition presented didactic instruction on dermatologic screening in video form, including images, classification schemes, and detailed explanations. All adaptive conditions robustly outperformed the non-adaptive control in both learning efficiency and fluency (large effect sizes). Between adaptive conditions, SDT retirement criteria produced greater learning efficiency than standard, accuracy-based mastery criteria at both immediate and delayed posttests (medium effect sizes). SDT sequencing and standard adaptive sequencing did not differ. SDT enhancements to adaptive perceptual learning procedures have potential to enhance learning efficiency.

Keywords: perceptual learning; adaptive learning; signal detection; medical image perception; skin cancer, dermatology, cancer image interpretation

Introduction

Skin cancer is the most common cancer in the US (CDC, 2018), with melanoma being the most lethal form of skin cancer. The number of adults treated for all skin cancers annually in the US grew from 3.4 million in 2002–2006 to 4.9 million in 2007–2011. The annual cost of treating all skin cancers in the US grew from \$3.6 billion to \$8.1 billion over this same period, with non-melanoma skin cancer (NMSC) treatment costs estimated at \$4.8 billion and the average annual cost of treating melanoma estimated at \$3.3 billion (Guy, Machlin, Ekwueme & Yabroff, 2015). The estimated incidence of melanoma for 2019 was 96,480 cases and 7,230 deaths (Siegel, Miller & Jemal, 2019). Additional costs attributable to melanoma and NMSC are estimated at \$39.2 million and \$28.9 million, respectively, for morbidity (e.g., lost income from being able to work or perform normal chores), and \$3.3 billion and \$1.0 billion for mortality (e.g., lost income due to premature death) (Guy & Ekwueme, 2011).

In dermatology and other medical domains, such as mammography and pathology, saving or extending lives often depends on effective visual detection and interpretation of suspicious findings in medical images. The tasks are challenging, and the relevant skills are complex. Recent estimates are that approximately 30% of potentially detectable cancers in cancer images are missed (Krupinski, 2010). Although artificial intelligence approaches, especially deep learning, have shown promise in skin cancer diagnosis (Zhou et al., 2021; Hekler et al., 2019), characteristic limitations, including algorithmic bias and effectiveness that varies with skin color (Chan et al., 2020), imply diagnostic decisions are likely to remain in human hands (and eyes) for the foreseeable future.

Research in recent years has shown the key role of *perceptual learning* (PL) in the development of expertise in medical, STEM, and other domains (for reviews, see Kellman & Garrigan, 2009; Kellman et al., 2022). Efforts to apply PL concepts have led to an emerging learning technology of PL that can accelerate pattern recognition, fluency, and transfer (Kellman et al., 2022; Kellman, Massey & Son, 2010). Often, strong and lasting PL effects can be produced by relatively short interventions.

Adaptive Perceptual Learning

Particularly fruitful in the development of useful PL technology has been the combination of perceptual learning concepts with adaptive learning methods (Kellman, Jacoby, Massey & Krasne, 2022). Perceptual-adaptive learning modules (PALMs) integrate an adaptive learning algorithm, Adaptive Response-time Based Sequencing (ARTS), into perceptual category learning. Trials in ARTS are interactive episodes in which one or more displays are presented and the learner must make an active response. Applied to learning perceptual classifications, each learning category is assigned a priority score indicating the relative benefit of a

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

novel instance of that category appearing on the next trial. The priority score for each category, updated after every trial, is a function of learner accuracy, response times, and trials elapsed since last presentation. (See Mettler, Massey & Kellman, 2011; Mettler, Massey & Kellman, 2016; and the current method section for computational details.) Efficiency is gained in ARTS because these properties of the algorithm make spacing sensitive to the interaction of each learner with each learning category. ARTS tends to optimize learning for all categories concurrently, and learning occurs with relatively few errors. ARTS also combines accuracy and response times to adaptively implement objective *mastery* criteria, typically defined as accurate responses to successive, spaced exemplars of a category in less than some criterion response time (RT). Retirement (dropout) of mastered categories is used to focus learner effort where it is most needed.

In medical learning, PALMS have been consistently shown to produce rapid acceleration in learning and long-lasting gains from relatively short interventions. These effects have been found across a wide variety of domains, including dermatology, histopathology, echocardiography, and ophthalmology (Evered, 2018; Romito et al., 2016; Krasne, et al., 2013; Rimoin et al., 2015; Kellman, et al., 2022). A *Basic Dermatology PALM* targeting the learning of 12 basic morphological categories in dermatological classification along with their secondary configuration and anatomical distribution, substantially improved medical students' diagnostic classification, and the improvement was maintained in assessments given one year later (Rimoin, Altieri, Craft, Krasne & Kellman, 2015). The PALM produced these results with training times averaging less than 30 minutes.

Signal Detection Theory and Adaptive Learning

From their inception 50 years ago (e.g., Atkinson, 1972) to the present, adaptive learning systems have uniformly used learner accuracy as a basis (usually, the only basis) for adjusting spacing or content. We believe that this reliance contains an important limitation, one that becomes especially salient in applications to PL of multiple categories. Using accuracy in category responses does not separate *sensitivity* in the signal detection sense from *criterion* or bias. Suppose a category in a multi-category Dermatology Diagnostic PALM is squamous cell carcinoma. Consider an observer who, in a training session, reports "squamous cell carcinoma" on each presentation of every image, from any category. All adaptive systems we know of would evaluate the learner as having learned the squamous cell carcinoma category, with this assessment based on trials on which instances of squamous cell carcinoma were actually presented. In SDT terms, when the learner is just as likely to say "squamous cell carcinoma" to squamous cell carcinoma images as to images of other categories, hit and false alarm rates are equal, and the observer actually has zero SDT sensitivity for squamous cell carcinoma.[1] An adaptive learning system such as ARTS (Mettler & Kellman, 2014) would eventually correct this because the learner would ultimately have to demonstrate mastery of competing or confusable categories. It is possible, however, that a more direct use of false alarm information might be more efficient.

Although SDT concepts have been used in radiological diagnosis and other medical domains to characterize outcomes or assess performance, these concepts have not been used as dynamic inputs to adaptive learning in medical training or elsewhere. Adaptive perceptual learning, and adaptive learning systems in general, might be improved by using concepts borrowed from signal detection theory (SDT).

At first glance, it would seem difficult to mesh adaptive learning and SDT concepts. SDT measures do not technically apply to learning, as they formally measure stationary quantities (i.e., determining sensitivity for a single signal strength with the assumption of unchanging noise and signal+noise distributions (Swets, 1979; Wickens, 2002). Studies of PL show that learning changes SDT sensitivity, by means of both *signal enhancement* and *noise suppression* (e.g., Dosher & Lu, 1999; Gold, Bennett & Sekuler, 1999). In other words, PL moves or alters the underlying distributions. Another apparent problem relates to estimating performance for categories in PL. Each learning trial for actinic keratosis, for example, will typically involve a novel example. In its pure form, SDT utilizes many responses to estimate the noise and signal+noise distributions for a single stimulus value. Within any learning category, however, instances will differ somewhat in their difficulty (signal strength). Another limitation in adaptive learning is that relatively large numbers of trials are typically needed to estimate formal SDT parameters.

Our starting point is that use of accuracy data in adaptive learning has an SDT-style problem in failing to distinguish criterion from sensitivity. Although formal estimation of SDT measures in ongoing adaptive learning is problematic, we can recast adaptive learning systems to incorporate SDT-style constraints. This has become common in psychophysical work where performance on a category having different exemplars is characterized in terms of an SDT sensitivity measure such as d′. The goal is not to determine an exact S+N distribution but to compare performance across time or conditions using a measure that disentangles sensitivity and response bias.

[1] For clarity, we refer to the notion of sensitivity from SDT as *SDT sensitivity*, to distinguish it from the notion of sensitivity of a medical test (referring to accuracy for positive cases). SDT sensitivity is an overall measure of detection performance (often measured by d' or area under an ROC curve), monotonically related to the difference between the hit rate and the false alarm rate. A medical test that has high diagnostic sensitivity but low specificity (high false-alarms) could have zero SDT sensitivity.

In the present work, we applied and tested SDT-style concepts to adaptive learning in two ways. First, we used it to alter category *spacing* during learning. In ARTS, applied to PL, category recurrence depends on the learner's accuracy and RT. An error in classifying a category exemplar causes early recurrence of that category (for details, see Mettler & Kellman, 2014). In forced-choice, multi-category classification, however, each *miss* for a presented category also comprises a *false alarm* for another category, and no current adaptive systems use that information for sequencing. For SDT-style sequencing, we treated the false-alarmed category as an error, mandating its rapid re-occurrence.

A separate SDT modification involved category *retirement* (dropout). Typically, ARTS uses learning to criteria, such that a category is retired when certain accuracy and RT benchmarks are met across several widely-spaced learning trials, and the learning phase ends when all categories have been retired. To utilize SDT concepts more directly, we implemented a different retirement scheme based on calculation of a running sensitivity (d') measure, applied to each learning category.

These modifications were incorporated into learning conditions in a Skin Cancer PALM that trained learners to classify skin lesions. We compared spacing and retirement in a baseline adaptive learning system (standard ARTS sequencing and retirement, an adaptive control) to modified adaptive systems that incorporated SDT-style spacing and retirement.

Method

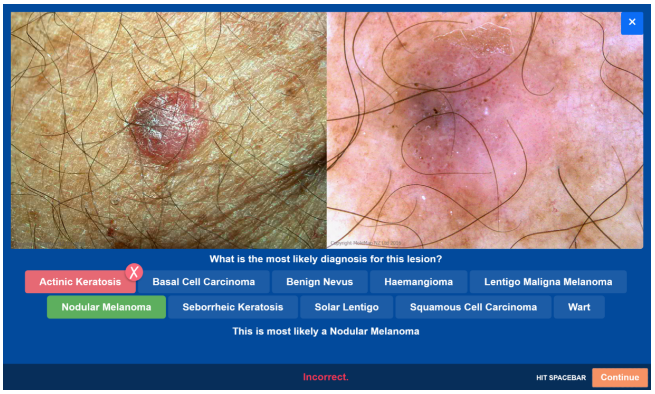
Participants

122 undergraduate UCLA students participated in person for course credit. Participants had no particular medical background or training.

Materials

The skin cancer PALM trained classification of 10 categories of cancerous and benign skin lesions. Cancerous categories included basal cell carcinoma, lentigo maligna melanoma, nodular melanoma and squamous cell carcinoma. Benign categories included actinic keratosis, benign nevus, haemangioma, seborrheic keratosis, solar lentigo, and wart. For each category we obtained between 18-110 individual exemplars. Each exemplar had two associated images, a clinical (macroscopic) view of the skin lesion and a dermoscopic image. A dermoscope incorporates high magnification and an adjustable illumination system that allows detailed assessment beneath the outer surface of the skin. Images were selected from a MoleMap, Inc. database based on: original dermatologic diagnosis, verification via biopsy when appropriate, and good image quality. Image diagnoses were verified by multiple sources including dermatologists and AI methods.

The PALM presented the paired macroscopic and dermoscopic images above choice labels for each of the 10 categories (see Figure 1). On each trial, a category exemplar was presented by showing the two images simultaneously, with the macroscopic image on the left and the dermoscopic image on the right. Images were 3200 x 1200 pixels in size and displayed to fill the available screen area resulting in an image 1888x708 pixels in size.



What is the most likely diagnosis for this lesion?

This is most likely a Nodular Melanoma

Incorrect.

Figure 1: Example of a PALM learning trial with feedback.

Trials in the learning and assessment phases were identical with the exception that no feedback was given in the latter. In the learning phase, feedback indicated the correctness of each response, highlighted the correct answer, and indicated how quickly the learner had responded. Half of assessment items could have appeared in the training set (although not every participant would see all training images) and the other half never appeared in training.

We tested 5 conditions: 4 adaptive conditions and one traditional instruction control. One adaptive condition was the standard ARTS algorithm used in prior studies, serving here as an adaptive control. The traditional instruction control condition consisted of a video-based traditional didactic presentation in 3 separate videos totaling 50 minutes long. It covered topics from lecture-based lessons on skin lesion classification.[2] There was sufficient visual information in the control condition videos to accurately classify exemplars from each category. In between each of the 3 videos was a short quiz on the information from the just-seen video, for a total of 2 quizzes.

Design

In the 4 adaptive conditions, a 2x2 design resulted from using all combinations of standard ARTS or SDT spacing and standard ARTS or SDT retirement. The four conditions were: ARTS Sequencing & ARTS Retirement; ARTS Sequencing & SDT Retirement; SDT Sequencing & ARTS Retirement; SDT Sequencing & SDT Retirement. The ARTS sequencing and ARTS retirement condition served as an adaptive control.

[2] Topics included: Introduction to histopathology, definitions, illustrations of skin cancer, categories of lesions, ABCDE criteria for diagnosis, dermoscopy and dermoscopic structures, vascular patterns and examples of lesions, among other related topics.

Sequencing

Sequencing refers to the algorithm determining the adaptive presentation of learning categories. ARTS sequencing was based on category priority scores for each trial based on accuracy, response time, and elapsed trials. The priority of each category on each trial was determined by the ARTS priority score equation shown in equation 1 and as described in (Mettler, Massey & Kellman, 2011; 2016).[3]

$$P_i = a(N_i - D)[b(1 - \alpha_i)\, Log(RT_i / r) + \alpha_i W] \qquad (1)$$

SDT-based sequencing utilized standard ARTS sequencing but added to it an enhanced priority for false-alarmed categories. A false-alarmed category would recur about 2-3 trials (± random jitter) later. The incorrectly answered category would also recur with high priority, subject only to a similar enforced delay. False-alarmed category recurrences were ignored for purposes of calculating retirement.

Retirement

Retirement refers to removal of a category from ongoing learning trials when performance on that category has reached mastery criteria. In ARTS retirement, as in previous studies, the learning criteria included accuracy and response times (RTs) for the last several presentations of the category. Items were retired when categories were responded to correctly in 5 of the last 5 presentations of the category and when each of those presentations had a RT of less than 15 seconds. Retired categories could still be presented after retirement in the form of filler presentations to create adequate trial spacing for unretired categories when the number of available unretired items was low.

The SDT retirement condition was based on a running SDT-style sensitivity measure approximating d-prime (d′). D-prime was calculated by subtracting a normal distribution transformed value of false alarms from a normal distribution transformed value of hits. A log-linear correction was used for cases of zero or perfect accuracy (Hautus, 1995). The running d’ measure was obtained by considering a window of hits and false alarms spanning the last 8 presentations of the target category. Hits were calculated as the proportion correct of target category presentations and false alarms were calculated using the proportion of target category false alarms out of the total number of possible occasions for false alarms – all non-target category presentations between the first instance of the target category and the last instance in the 8 presentation window. The category was retired if the resulting d′ value was above 2.4. Retirement was calculated and implemented after each target category presentation, and was not calculated or implemented on non-target category trials. Due to a programming error, some categories in the first two thirds of the experiment were not retired at the correct d′ value and were instead retired slightly before the intended d′ value. Besides the sensitivity criterion, the SDT retirement condition included the same RT criterion as in the standard ARTS retirement condition.

Procedure

There were 2 experimental sessions administered one week apart. Session 1 consisted of a pretest, learning phase (either non-adaptive control condition or an adaptive condition) and immediate posttest. Session 2 consisted of a delayed posttest 7 days after Session 1.

Participants were randomly assigned to one of five conditions (20 participants per condition after participant exclusions). A pretest balancing algorithm ensured that average pretest accuracies were comparable across conditions. After pretest each participant was assigned to the condition that would best equate the running average pretest accuracy for prior subjects (see Mettler, Massey, Burke, Garrigan & Kellman, 2018); this procedure is similar to minimization (Pocock & Simon, 1975)

The pretest, posttest and delayed posttest for all conditions were conducted in the PALM trial format, but without feedback. Non-adaptive control participants were instructed to watch the video lecture without skipping or rewinding. Adaptive condition participants were instructed to read the on-screen instructions and continue through the PALM at their own pace. In assessment and adaptive learning, participants were given 30 seconds to respond on each trial.

Dependent Measures and Data Analyses

There were 2 assessment versions distributed in two orders across the 3 test phases. If one version was administered as a pretest, the alternate version was administered as an immediate posttest and the initial version was administered again as the delayed posttest. The assessment orders were randomly assigned to participants and balanced across experimental conditions.

There were two assessment items for each category except for Benign Nevus, which was split into two subcategories, Compound and Junctional, with 2 items each, for a total of 22 items on each assessment. Due to an error, one assessment item in the Lentigo Maligna Melanoma category was incorrectly assigned, and was omitted from analyses, making the total number of assessment items 21. In each category, one assessment item had been available during training, and the other item was a novel item never shown during training. The order of assessment items in each version was randomized and fixed, with the constraint that the same category did not appear in adjacent trials more than once in a given assessment.

Exclusion criteria Participants were excluded from the study if they achieved greater than 30% accuracy on the pretest. The exclusion occurred immediately after the pretest, and such participants (20) did not complete the learning phase. Participants were excluded after data

[3] ARTS parameters for the study included: Enforced delay, D = 3 trials (plus or minus jitter of .5); RT weight, r= 4; Default priority weight = 1; Target retirement RT = 15 seconds; Timeout = 30 seconds. Other parameters were similar to prior studies (Incorrect penalty W = 20; weighting constants: a = .1, b=1.1).

collection if they spent more than 700 trials in training or if accuracy for the entire learning phase was less than 30%. Two participants were excluded on this basis.

Dependent Measures Our primary measure of learning performance was *learning efficiency*. Learning efficiency is posttest accuracy divided by the time (or number of trials) invested in learning. Efficiency is useful when mastery criteria are used. Participants will take differing amounts of time to reach mastery, and final performance is likely to be similar in conditions with similar or the same mastery criteria. Efficiency comprises a single measure that combines two types of performance data – learning time invested and posttest accuracy – into one rate measure. Accuracies, response times, and time or trials to criterion were also examined separately.

For most dependent measures, both a one-way and factorial ANOVA were conducted. The factorial design included two factors, Sequencing (ARTS vs. SDT) and Retirement (ARTS vs. SDT) that applied to each adaptive condition. However, because neither factor correctly applied to the control condition, a separate, one-way ANOVA was conducted on all conditions for each dependent measure. In addition to factors for sequencing and retirement, two additional factors were included, assessment version and posttest phase. A final factor, whether assessment items were seen or not seen during training, was excluded from the analyses because, in practice and depending upon performance, not all seen items on the assessment were seen by participants in the training phase. Collapsing across seen and unseen items simplified interpretation of the analyses.

All graphs show means +/– one standard error of the mean. Condition names are sometimes abbreviated as follows: AA: ARTS Sequencing and ARTS Retirement, AS: ARTS Sequencing and SDT Retirement, SA: SDT Sequencing and ARTS retirement, SS: SDT Sequencing and SDT retirement, and C: control.

Results

Learning Efficiency

Learning efficiency was examined by both time and trials. Efficiency by time was accuracy gain from pretest to posttests divided by total time spent in the learning phase. Efficiency by trials divided gain by the number of learning trials invested. Only efficiency by time could be calculated for the traditional instruction control condition since it did not contain learning trials.

Efficiency by Time Efficiency by time is shown in Figure 2. Efficiency by time was noticeably higher for all adaptive conditions vs. the control condition. Adaptive conditions with SDT retirement had higher efficiencies than adaptive conditions with standard ARTS retirement. These observations were confirmed by analyses. A 5x2x2 ANOVA on Condition, Posttest Phase, and Assessment Version was carried out. There was no main effect or interaction involving Assessment Version in this or any other analysis, so results for this factor have been omitted below. We found a significant effect of Condition (F(4,90)=13.977, p<.001, η_p^2 =0.383), a significant effect of Posttest Phase (F(1,90)=36.797, p<.001, η_p^2=0.29), and a reliable interaction of Condition and Posttest Phase. Paired comparisons were conducted between all conditions. All paired comparisons between adaptive conditions and the conventional instruction control condition *were highly significant with large effect sizes* at both posttests: at immediate posttest, all ps<.001, Cohen's Ds ranged from 2.24-2.53; at delayed posttest, all ps<.001, Ds ranged from 2.32-2.68. Between adaptive conditions, the following paired comparisons were significant or marginally (p<.10) significant: At immediate posttest, AS vs. AA (t(38)=2.061, p=.046, D=0.69), AS vs. SA (t(38)=1.70, p=.098, D=0.55); at delayed test, SS vs. SA (t(38)=1.82, p=.077, D=0.58). The interaction between Condition and Test Phase was likely driven by reliable changes in efficiency between immediate and delayed posttest for all adaptive conditions (all ps<.02), but marginal change for the non-adaptive control condition, (t(19)=1.93, p=.069, D=0.518).

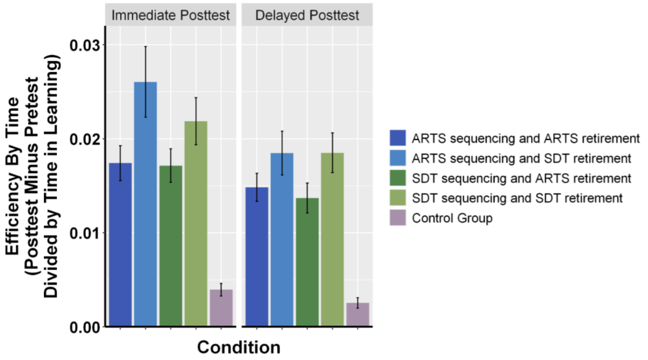


Figure 2: Efficiency (time-based) by Condition.

The SDT and adaptive control conditions were compared in a separate 2x2x2x2 ANOVA with Sequencing, Retirement, Posttest Phase and Assessment Version as factors. SDT retirement was significantly superior to ARTS retirement, as shown by a main effect of Retirement (F(1,72)=4.861, p=.031, η_p^2 =0.063), and this effect was consistent across posttest phase (no reliable interaction of Retirement and Phase, p = .115). There was a significant main effect of Posttest Phase (F(1,72)=30.357, p<.001, η_p^2=0.297), but no reliable effect of Sequencing (F(1,72)=0.44, p=.509, η_p^2=0.006). Paired comparisons across levels of each factor showed a significant difference between ARTS and SDT retirement (t(78)=2.55, p=.013, D=0.585), no significant difference between ARTS and SDT sequencing (t(78)=0.625, p=.534, D=0.14), and a significant difference between immediate and delayed posttest (paired t-test: t(79)=5.61, p<.001, D=0.409).

Efficiency by Trials Efficiency calculated by trials was similar to efficiency by time; the primary result was that SDT retirement was superior to ARTS retirement, p=.012.

Accuracy Gain Accuracy gain for all five conditions is shown in Figure 3. All adaptive conditions had higher accuracy gain than the control condition. Accuracy gains

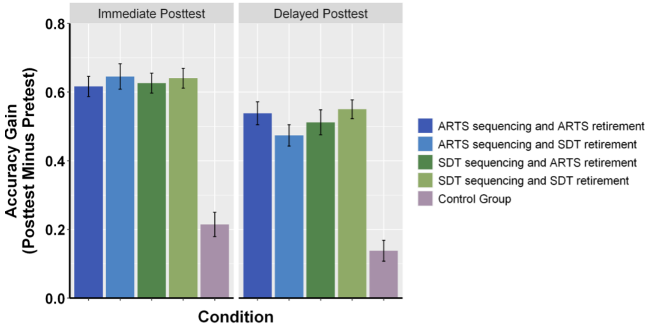


Figure 3: Accuracy Gain by Condition.

were highly similar across adaptive conditions at immediate posttest and showed modest numerical differences at delayed posttest. A 5x2x2 ANOVA on Condition, Posttest Phase, and Assessment Version found significant effects of Condition (F(4,90)=44.098, p<.001, η_p^2=0.662) and Posttest Phase (F(1,90)=45.11, p<.001, η_p^2=0.334), and no reliable interactions. Paired comparisons were conducted between all adaptive conditions and the conventional instruction control condition. All t-tests were highly significant: At immediate posttest, all ps<.001, Ds ranged from 2.66-2.97; at delayed posttest all ps<.001, Ds ranged from 2.45-3.16.

A separate ANOVA conducted on only the adaptive conditions showed no significant effect of Sequencing (F(1,72)=0.164, p=.686, η_p^2=0.002) or Retirement (F(1,72)<0.001, p=.984, η_p^2<0.001), and a significant effect of Phase (F(1,72)=46.01, p<.001, η_p^2=0.39).

Duration of Learning Trials to retirement (A) and learning duration by time (B) are shown in Figure 4.

Duration of learning was computed separately by trials and by time, primarily because the conventional instruction control condition did not have active learning trials (only 17 quiz question trials).

For duration by trials, a 2x2 ANOVA was conducted on Sequencing and Retirement. The ANOVA found no significant main effect of Sequencing (F(1,76)=0.67, p=.416, η_p^2=0.009), a significant main effect of Retirement (F(1,76)=12.68, p<.001, η_p^2=0.143) and no significant interaction (F(1,76)=2.45, p=.122, η_p^2=0.031).

Paired comparisons conducted between each adaptive condition using learning duration in trials found reliably shorter durations for SS vs. AA (t(38)=2.52, p=.016, D=0.815); SA vs. AS (t(38)=2.295, p=.027, D=0.727); and SS over SA (t(38)=3.90, p<.001, D=1.31).

For duration by time, a 5 way ANOVA was conducted on each condition. There was a significant main effect of condition (F(4,90)=6.09, p<.001, η_p^2=0.213), due to longer duration of learning in the non-adaptive condition. A second 2x2 ANOVA was conducted on only the adaptive conditions using Sequencing and Retirement as factors. The ANOVA found no significant effect of Sequencing (F(1,76)=0.781, p=.380, η_p^2=0.01), no significant effect of Retirement (F(1,76)=2.584, p=.112, η_p^2=0.033) and no significant interaction (F(1,76)=0.034, p=.854, η_p^2<0.001). Paired comparisons were conducted between each adaptive condition using learning duration in minutes. T-tests found a marginally shorter learning duration for AS vs. SA (t(38)=1.696, p=.098, D=0.538). No other comparisons were significant, all ps > .128).

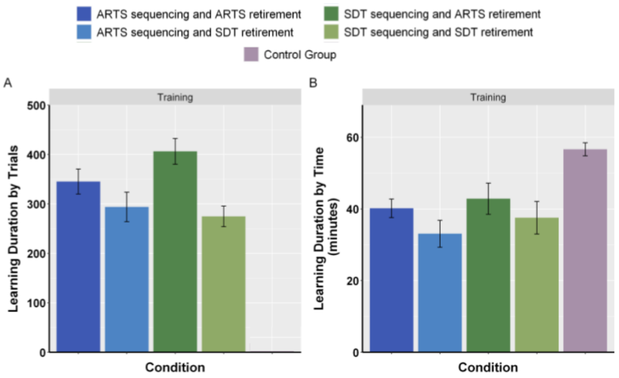


Figure 4: A. Duration of learning in trials (i.e., trials to retirement) and B. Duration of Learning in Minutes (i.e., minutes to retirement).

Discussion

We tested adaptive perceptual learning in skin cancer screening with and without SDT modifications and compared the adaptive conditions to conventional instruction that included examples and standard guidelines for assessing and classifying cancerous and non-cancerous skin lesions. All of the adaptive perceptual learning conditions robustly outperformed the control group, with very large effect sizes. These results reinforce the emerging conclusion that instructional interventions designed to advance PL address crucial components of medical learning that are not well addressed by conventional instruction (Kellman & Massey, 2013; Kellman et al., 2022).

In four adaptive conditions, we tested two new schemes based on signal detection principles: SDT-based sequencing, which specifically incorporated direct consequences of false alarms into the flow of learning events, and SDT-based retirement. These modifications were compared to standard ARTS which served as an adaptive control. We found that SDT-based retirement led to markedly improved efficiency relative to standard, accuracy-based retirement. In contrast, we did not find significant differences between standard ARTS sequencing and SDT-based sequencing. In SDT-based retirement, utilizing a running d′ measure of sensitivity to assess mastery produced equivalent accuracy with a smaller investment of learning, with moderate effect sizes at both immediate and delayed posttests. By incorporating both hit rate and false alarm rates and basing learning criteria on d′, SDT-based retirement provides a more discerning method of assessing perceptual category mastery.

Future investigations will likely profit from further exploring the space of possible SDT-based enhancements to adaptive perceptual learning.

Acknowledgements

We gratefully acknowledge support for this research from National Institutes of Health grant R01CA236791.

For programming, experiment design and assistance running participants, we thank Tim Burke, and members of the UCLA Human Perception Lab, Faye Tesoro, Mindy Kim, Amy van Beek, Wonjoon Choi, Alexandra Kostrikova, Brandon Achugbue, Hoi Yu Blair Chong, Elisa Fang, Grace Kim, Claudia Chen, Arianna Chow, Nathan Wang and Faith Mackriss.

References

Atkinson, R. C. (1972). Optimizing the learning of a second-language vocabulary. *Journal of Experimental Psychology*, 96(1),124–129. http://dx.doi.org/10.1037/h0033475

CDC – Skin Cancer Statistics. (2018, May 29). Retrieved May 10, 2019, from https://www.cdc.gov/cancer/skin/statistics/index.htm

Chan, S., Reddy, V., Myers, B., Thibodeaux, Q., Brownstone, N., & Liao, W. (2020). Machine learning in dermatology: Current applications, opportunities, and limitations. *Dermatology and Therapy, 10*(3), 365–386. https://doi.org/10.1007/s13555-020-00372-0

Dosher, B. A., & Lu, Z. L. (1999). Mechanisms of perceptual learning. *Vision Research, 39*(19), 3197–3221. https://doi.org/10.1016/S0042-6989(99)00059-0

Evered, A. (2018). Perceptual and adaptive learning modules and their potential to transform cytology training programmes. *Cytopathology, 29*(4), 371–374. https://doi.org/10.1111/cyt.12578

Gold, J., Bennett, P. J., & Sekuler, A. B. (1999). Signal but not noise changes with perceptual learning. *Nature*, 402(6758), 176–178. https://doi.org/10.1038/46027

Guy, G. P., & Ekwueme, D. U. (2011). Years of potential life lost and indirect costs of melanoma and non-melanoma skin cancer: A systematic review of the literature. *PharmacoEconomics*, 29(10), 863–874. https://doi.org/10.2165/11589300

Guy, G. P., Machlin, S. R., Ekwueme, D. U., & Yabroff, K. R. (2015). Prevalence and costs of skin cancer treatment in the U.S., 2002–2006 and 2007–2011. *American Journal of Preventive Medicine*, 48(2), 183–187. https://doi.org/10.1016/j.amepre.2014.08.036

Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values ofd′. Behavior Research Methods, Instruments, & Computers, 27(1), 46–51. https://doi.org/10.3758/BF03203619

Hekler, A., Utikal, J. S., Enk, A. H., Hauschild, A., Weichenthal, M., Maron, R. C., Berking, C., Haferkamp, S., Klode, J., Schadendorf, D., Schilling, B., Holland-Letz, T., Izar, B., von Kalle, C., Fröhling, S., Brinker, T. J. (2019). Superior skin cancer classification by the combination of human and artificial intelligence. European Journal of Cancer, 120, 114–121. https://doi.org/10.1016/j.ejca.2019.07.019

Kellman, P. J. (2002). Perceptual learning. In R. Gallistel & H. Pashler (Ed.), *Stevens' Handbook of Experimental Psychology, Third Edition, Vol. 3* (Learning, motivation and emotion), John Wiley & Sons.

Kellman, P. J. (2013). Adaptive and perceptual learning technologies in medical education and training. *Military Medicine, 178*(10S), 98–106. https://doi.org/10.7205/MILMED-D-13-00218

Kellman, P. J., & Garrigan, P. (2009). Perceptual learning and human expertise. *Physics of Life Reviews, 6*(2), 53–84. https://doi.org/10.1016/j.plrev.2008.12.001

Kellman, P. J., Jacoby, V., Massey, C., Krasne, S. (2022). Perceptual learning, adaptive learning, and gamification: Educational technologies for pattern recognition, problem solving, and knowledge retention in medical learning. In: Witchel, H.J., Lee, M.W. (eds) *Technologies in Biomedical and Life Sciences Education. Methods in Physiology.* (pp. 135-166). Springer, Cham. https://doi.org/10.1007/978-3-030-95633-2_5

Kellman, P. J., & Krasne, S. (2018). Accelerating expertise: Perceptual and adaptive learning technology in medical learning. *Medical Teacher, 40*(8), 797-802. doi: https://doi.org/10.1080/0142159X.2018.1484897

Kellman, P. J., & Massey, C. M. (2013). Perceptual learning, cognition, and expertise. In B. H. Ross (Ed.), *Psychology of Learning and Motivation (Vol. 58*, pp. 117–165). Academic Press. https://doi.org/10.1016/B978-0-12-407237-4.00004-9

Kellman, P. J., Massey, C. M., & Son, J. Y. (2010). Perceptual learning modules in mathematics: Enhancing students' pattern recognition, structure extraction, and fluency. *Topics in Cognitive Science, 2*(2), 285–305. https://doi.org/10.1111/j.1756-8765.2009.01053.x PMCID: PMC6124488

Krasne, S., Hillman, J. D., Kellman, P. J., & Drake, T. A. (2013). Applying perceptual and adaptive learning techniques for teaching introductory histopathology. *Journal of Pathology Informatics*, 4(1), 34. https://doi.org/10.4103/2153-3539.123991

Krupinski, E. A. (2010). Current perspectives in medical image perception. *Attention, Perception & Psychophysics*, *72*(5), 1205–1217. https://doi.org/10.3758/APP.72.5.1205

Mettler, E., & Kellman, P. J (2014). Adaptive response-time-based category sequencing in perceptual learning. Vision Research, 99, 111–123. https://doi.org/10.1016/j.visres.2013.12.009 PMCID: PMC6124487

Mettler, E., Massey, C. M., Burke, T., Garrigan, P., & Kellman, P. J. (2018). Enhancing adaptive learning through strategic scheduling of passive and active learning modes. In T.T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), Proceedings of the 40th Annual Conference of the Cognitive Science Society (pp. 768-773). Austin, TX: Cognitive Science Society.

Mettler, E., Massey, C.M., & Kellman, P. J. (2011). Improving adaptive learning technology through the use of response times. In L. Carlson, C. Hölscher, & T.

Shipley (Eds)., *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society*. Boston, MA: Cognitive Science Society, 2532-2537.

Mettler, E., Massey, C. M., & Kellman, P. J. (2016). A comparison of adaptive and fixed schedules of practice. *Journal of Experimental Psychology: General*, 145(7), 897–917. https://dx.doi.org/10.1037/xge0000170 PMCID: PMC6028005

Pocock, S. J., & Simon, R. (1975). Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. Biometrics, 31(1), 103–115.

Rimoin, L., Altieri, L., Craft, N., Krasne, S. & Kellman, P. J. (2015). Training pattern recognition of skin lesion morphology, configuration, and distribution. Journal of the American Academy of Dermatology, 72(3), 489–495. https://doi.org/10.1016/j.jaad.2014.11.016

Romito, B. T., Krasne, S., Kellman, P. J., & Dhillon, A. (2016). The impact of a perceptual and adaptive learning module on transoesophageal echocardiography interpretation by anaesthesiology residents. British Journal of Anaesthesia, 117(4), 477–481. https://doi.org/10.1093/bja/aew295

Siegel, R. L., Miller, K. D., & Jemal, A. (2019). Cancer statistics, 2019. CA: A Cancer Journal for Clinicians, 69(1), 7–34. https://doi.org/10.3322/caac.21551

Swets, J. A. (1979). Roc analysis applied to the evaluation of medical imaging techniques: Investigative Radiology, 14(2), 109–121. https://doi.org/10.1097/00004424-197903000-00002

Wickens, T. D. (2002). *Elementary Signal Detection Theory*. Oxford University Press.

Zhou, S. K., Greenspan, H., Davatzikos, C., Duncan, J. S., Van Ginneken, B., Madabhushi, A., Prince, J. L., Rueckert, D., & Summers, R. M. (2021). A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proceedings of the IEEE*, 109(5), 820–838. https://doi.org/10.1109/JPROC.2021.3054390