I. Introduction

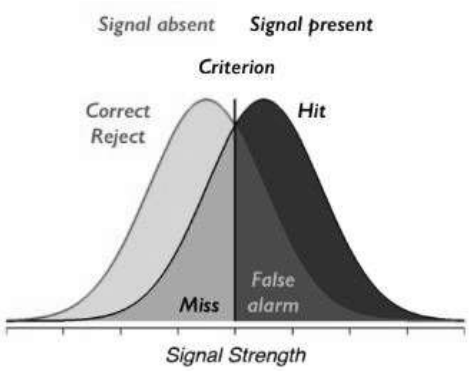
How much of cognition is perception? A large body of previous research suggests that much of expertise and mastery lies in *perceptual learning* (PL)—that is, when we see something and distinguish it enough times, we essentially learn to see that kind of thing better and pick up on higher level relations and concepts in those items the more we practice, and this is the true mark of an expert (Kellman & Massey 2013). Commonly held views of perception maintain that low level sensory information isn't important enough to have consequences for abstract thinking and learning, but current research hypothesizes that perception belies much more of learning than the traditional view suggests, and that learning could be improved in schools by adopting perceptual learning technologies rather than relying on the explicit rule-based teaching methods one might typically see in a classroom lecture (ie. memorizing facts and procedures).

Researchers in the field of perceptual learning champion the idea that you don't need to form an account in declarative knowledge in order to learn what things are, and that even when you do need a declarative or procedural account of steps to solve a problem, you still need to know how to apply the steps, when certain procedures are relevant, etc. (Kellman & Massey 2013). You need to be able to see what actually matters when applying the knowledge, and this understanding of structural patterns comes from seeing something enough times—a.k.a. perceptual learning.

The past decade of research at our lab has involved systematic efforts to study and apply PL technology in mathematics and science learning, specifically through the use of Perceptual / Adaptive Learning Modules (PALMs) that get learners to recognize patterns in structure and relations and to become more fluent in their processing of structure through repeated, varying interactions with instances of different kinds of stimuli. Considering the adaptive nature of perception—that we process things differently each time we look at them—recent efforts have adopted adaptive methods for sequencing (how much to space items) and retirement (when to stop showing an item that's learned) for item categories based on response time. Rather than just changing the presentation order when a user gets a question wrong, studies like (Mettler, Massey, and Kellman 2011), (Mettler and Kellman 2014), and (Mettler, Massey, Kellman 2016) employed adaptive learning algorithms that would space and retire categories based on how fast a participant could respond correctly, ultimately showing that such response-time-based category sequencing and retirement significantly increases efficiency of PL and enhances generalization to novel stimuli.

Our current study builds upon these developments in adaptive perceptual learning by introducing new adaptive learning schemes for sequencing and retirement based on the false alarm (FA) rate for categories along with accuracy, rather than on accuracy and response time alone. Using measures from the Signal Detection Theory (SDT) framework, we compare performance with these new adaptive approaches to performance under previous versions of the algorithm.

Signal Detection Theory, a statistical framework for understanding perceptual decision processes under conditions of uncertainty, is used to describe the sensitivity of an observer in detecting a signal against noise, or in our case, the sensitivity of an observer in discriminating between multiple noisy signals. This approach envisions two stimuli as statistical distributions of stimulus energy (like two normal distributions overlapping), that have a crucial decision 'criterion' point between them at which stimulus energy detected above the criterion level will be classified as from the higher of the two distributions, and stimulus energy below criterion will be classified as from the lower—so as that answers above the criterion will either be hits for the category with the higher distribution or false alarms for the category with the lower distribution, and vice versa.



From: http://gru.stanford.edu/lib/exe/fetch.php/tutorials/nobias.png?w=&h=&cache=cache

The distance between these distributions is essentially the user's degree of discriminability, which we call 'sensitivity' and measure with a value called D' or D-prime (mathematically, the z-score for hits minus the z-score for false alarms). 0 sensitivity implies no difference between hits and false alarms, and thus either no ability to detect differences between stimuli or no inherent difference in the stimuli themselves. Meanwhile, a value of 2 or 3 would imply that two distributions are 2 or 3 standard deviations apart.

By taking hits for a category and adjusting for the number of times that category was a false alarm for trials with a different target in some recent window, the D' value in SDT conditions takes into account not only our ability to recognize a category when it's in front of us, but also how much we might be confusing it for other categories. Reaching a certain D' criteria as a measure for retirement should ensure items aren't retired until they reach both sufficient levels of accuracy and sufficient lack of false alarms.

In each of 4 conditions, we manipulated either the sequencing or retirement scheme to have participants use either the tried-and-true ARTS system (accuracy and response time) or the new SDT system (accuracy and false alarms) to learn to correctly categorize types of skin cancer lesions. Thus, our conditions were: ARTS sequencing/ARTS retirement, ARTS sequencing/SDT retirement, SDT sequencing/ARTS retirement, and SDT sequencing/SDT retirement. Performance in these conditions was compared to a control group that received a traditional academic lecture on types of skin cancer.

Whereas in ARTS sequencing, only one category is 'punished', so to say, for an incorrect answer (that category is repeated soon after), in SDT sequencing, both the target category that was missed and the incorrect category choice will soon be shown again. And unlike ARTS retirement, which relies on the last 5 trials of a category all being answered correctly in under 10

seconds, SDT retirement requires that a category reach above a certain D' threshold (2.4) of discriminability (a certain amount of accuracy for a category - false alarms for it).

In theory, these SDT methods show great potential for correctively adapting the learning process to cover those areas a learner is actually most confused on. In practice, however, we find a much greater challenge in adapting signal detection/discrimination models to our 10-category alternative choice task, some aspects of which will be described below.

II. Reducing Noise in False Alarm Rates

The manner in which we're forced to calculate the false alarm rate of a category to get D' is tricky in our case, an issue we glossed over previously. For our D' calculation, we count the amount of hits among the last 8 trials, and we use however many nontarget trials were in that timeframe to count false alarms. An issue we immediately run into is that there is likely to be many more non-target trials than trials in this window, meaning the base rate of target occurence is a lot less than for non-targets, and that someone answering randomly would have way less hits than false alarms, ending up with a negative D' (when statistically, random answers should lead to a D' of 0, and a negative D' should imply purposefully incorrect responses).

Due to the way categories are adaptively sequenced, correct responses might lead to incredibly long windows for counting false alarms, which are also bound to contain less relevant material to what's confusing about that category the participant is getting right because the user isn't being shown items that they might have confused it with like they would be if they had answered the category incorrectly. Therefore, many of the categories participants run into during their false alarm window might not even really be possible confusions with the category we're trying to calculate a false alarm rate for.

The way we get around these issues now is by dividing the total number of false alarms across a window by the total number of categories to get some sensible measure of a false alarm rate that we can compare to a single category's hit rate. This approach makes some crude sense, and fixes our issue of getting a negative D', but poses problems of its own.

Low False Alarm Rates for Categories that Don't Present Often as Possible Confusions

For instance, we still don't have a way to enforce that a participant sees possible confusions, and if a category didn't have many opportunities to false alarm (possible confusions weren't shown often enough to accurately gauge FA rate), the false alarm rate might be drastically low, such that D' values are artificially inflated for those categories. When a participant is getting a category right, they have less opportunities to be wrong. It's hard to implement a change that solves this problem, because the point of our algorithm is to adapt based on what the user got wrong, not on what they *might* get wrong, but it may be helpful at this stage to just consider any possible circumstances that limit the opportunity for a user to false alarm with a category.

False Alarm Rates Consider Old, Irrelevant Signals

Another problem our false alarm averaging approach doesn't actually solve is the questionable relevance of the signals based on time observed. Because the adaptive sequencing might create huge windows for counting hits and false alarms, those earliest observed signals might have occurred, say, 100 trials before the most recent one and may represent the user's learned state much less than more recent signals would. Therefore, a user that took longer to retire a category in their 'final stretch' (perhaps by answering more correctly) had a less reliable measure of hits vs. false alarms by the end of it than one who retired a category based on a short span of trials (perhaps achieved by getting less false alarms but not necessarily answering as correctly). This phenomenon is hard to adjust for after the fact, as a longer window at retirement doesn't necessarily mean a participant is 'doing better'. One possible solution might be in actually weighing the strength of an individual hit or false alarm based on how recent it is to the user's current state of learning, so as we're not counting what's irrelevant.

False Alarm Rates Consider Irrelevant Categories

Finally, there's a qualitative issue with the D' data we get from averaging the amount of false alarms of a category like we do, in that we're not completely making sure all of the user's confusions have been settled by retirement, but more so hoping. In reality, every category should have a certain number of possible confusion categories that we use to divide the total false alarms by in order to get an accurate judgment of the sensitivity for that category with D', rather than assuming that a category could be equally confused among all the multiple alternatives (and once we figure that out, getting the participant to actually see all those categories within a retirement window is a separate issue). Getting this number of 'possibly confusable categories' might not be as straightforward as it sounds, but with the use of distance clustering algorithms, we might not be so far away from attaining a reliable, general continuous measure of 'confusability' across all the items in our study, which we could then use to infer how much these false alarm totals need to be divided up in order to get false alarm rates that accurately correspond to the problem space in discerning each category.

III. Adjusting the Criteria to Correct For Noise in False Alarm Rates

When we stop to consider that every category isn't as confusable as all the others, and seek to change the way we calculate false alarm rates, we're also begged the question: Why do we assume all categories should have the same criteria? After all, those instances of more easily confused categories obviously pose harder problems for discrimination than the most distinct instances. Moving the goalposts themselves might be an easier way to account for how variably learnable our categories are than changing the weight of false alarms based on how relevant the false alarms were to actual category confusions.

Pairwise Criteria

One way to address this problem could be by establishing pairwise D' criteria, such that every category has a unique goal to meet with every other category in terms of discriminability. This way, categories aren't being retired when they meet some standardized average level of

discriminability, but when we know discrimination has reached the standard set by the natural difficulty of discerning each possible pair of categories.

A huge problem with this idea though, outside of the mess of data it would create, is that, again, some categories might not have much of an opportunity to act as false alarms for other categories. If we're already worried about this creating exceptionally low false alarm rates in our current scheme (averaging total category false alarms in a window across the number of categories), then it's obvious that some of these pairwise criteria can't be reliably met.

We could possibly get around this issue by requiring that special pairwise D' criteria are only needed when a category has false alarmed for another a certain amount of times throughout the study. If two categories aren't being confused often, we just use the original D' criteria of 2.4. Therefore, we'd only help a learner by adjusting the criteria (either more or less learning) if the participant is having trouble—if the system isn't working, we get realistic about what they need to know. And, helpfully, if the user is having trouble with a category, they'll likely see more opportunities to false alarm with it, allowing us to get the accurate FA measures we need for pairwise criteria.

Different Criteria for Easy/Hard Categories

Another way we could approach this problem of mismatched category difficulty is by looking at how confusing a category is in general and moving its criterion based on that, rather than giving each category different goals with each other category.

For outliers, those categories that are least like the others and most recognizable, false alarms are likely less meaningful in terms of a miss because the category is less typical (though this may only be true of instances, as there's question to whether any of our categories are entirely outliers). The false alarm rate should likely be lower in actuality than what we see because participants might often know the proper typical distinctions between categories, but miss when they come across something they haven't been trained to see (and probably shouldn't be training for). Thus, participants may be overtraining on the easiest categories, and we should lower the required learning criteria for outliers.

For inlying categories, those that are closest together in resemblance, we can actually assume that the false positive rate is higher than what we measured, as many of the 'correct' answers were likely actual misses (but the participant accidentally got it right anyway since those categories are so similar). Therefore, participants may not be training enough on the hardest categories, and we should raise the criteria for inliers.

IV. Addressing Priorities in Learning

One might note that by making the criteria lower for easier categories, a participant could fly by the experiment hardly having learned those easy categories, but these issues could be further addressed by some kind of learning priority scheme in future studies, such as enforcing that the learner must absolutely learn the easiest categories, and that there will be even more focus on

those categories if they're struggling, as well as less emphasis on harder categories that we know are going to be hard.

Enforcing Easy/Hard Priorities in SDT Sequencing

One way we could potentially do this is by establishing a graded weighting scheme for affecting sequencing such that categories that are naturally easier will be repeated more often if the participant is confusing them and those that are harder won't be repeated as often even if the participant is confusing them (until the easier questions are retired).

This kind of graded punishment based on difficulty of misses in sequencing might work, but must be balanced with our main goal of still showing the categories a user got wrong more often and showing those they get right less often, which makes it complicated. For example, what happens when someone gets a hard category wrong? We'd have to subtract some 'confusability score' for how hard the category is from the fact that they still got it wrong, in terms of how soon we want to show the category again, and do the same for the category that was falsely alarmed. (Though, we don't necessarily have to show hard categories with net less frequency after a miss, just slightly less often than we would have if an easy category was missed).

Grading the sequencing punishment like this would allow a user to essentially focus on the easy stuff first before moving on to the harder problems, without totally separating easy and hard problems into separate phases of learning (though that could also be explored in future studies). There is some intuitive sense in this approach, in that it often feels like learning to solve easier problems prepares you for solving harder problems—perhaps the rules a learner builds when learning to distinguish the easier categories are built on when it comes time to distinguishing the harder ones, and by that point, they have more confidence in their abilities. In real life, though, doctors aren't going to see the 'less confusable' instances less often than they will see 'more confusable' instances–each variation of skin cancer will have its own level of prevalence in the population that doesn't correlate to a simple measure of confusability. If our goal is to train a doctor on the *experience* of distinguishing skin cancer lesions in the world, learning to do so in a more randomized setting (in terms of difficulty) like our current version of the study might have more long term effects—the learner may develop rules for handling uncertainties that they wouldn't have otherwise developed if they had only been tasked with telling apart relatively easy instances from one another and only relatively hard instances from one another.

Even more problems with this method come again from the issue of inadequate exposure to possible false alarms in the FA rate calculations for our easiest categories. By tackling the easy problems first then moving onto the harder problems, we'll likely ensure that we get more relevant trial groupings in the harder categories' nontarget trial windows for retirement, but easy categories will have seen mostly other easy, non-confusable instances in the nontarget trial window, meaning the FA rates for our easiest categories will become even less relevant to what you'd actually confuse those categories for. Prioritizing the learning of easy categories in SDT sequencing may prove effective for the learning process, but we must also make sure the learner can discriminate between the easy and hard categories as well at the end of the day.

V. Need for a General Recognition Model

Most of the issues discussed so far point to a more glaring issue in our study in that we don't actually have a model for general recognition that takes into account the multitude of options available to the user in this task. Though we've been able to adapt measures from the single-class signal detection model first described on page 2 to analyze performance in our SDT conditions, we've had to work around the fact that the model wasn't built to handle the level of uncertainty in responses that we have in our trials.

Whereas in single-class signal detection, you can always attribute a hit for the signal stimulus to a miss for the noise stimulus and a miss for the signal to a false alarm for the noise, in our tasks, a hit or miss for one category signal might entail multiple false alarms for other category signals (maybe the user thought an instance of Actinic Keratosis looked 35% like Solar Lentigo, 25% like Benign Nevus, and 40% like Actinic Keratosis). In fact, maybe a miss doesn't equate to any false alarm at all if the user was completely guessing. Incorrect answers could imply actual false alarms or just pure guesses without a category in mind, and correct answers could be actual true positives or just guesses that were accidentally right. We're never sure where the scale lies in terms of how much of any answer is a guess vs. how much of that answer represents belief in a choice, and we're not sure what other choices the user might believe in at the same time—both of which would inform us greatly on the user's actual learned state.

Confidence Ratings Can Deduce the Amount of Guess in Responses

One way to intuit the ratio of guess-to-answer could be to include confidence ratings with each trial response. By combining each answer with an indication of the participant's confidence that the response is correct, we get evidence about where on the underlying decision axis the effect of the stimulus falls (Wickens). Creating even just three levels of confidence (like 'guessing', 'uncertain', 'sure') for each answer allows us to capture much more data from the single-class detection model, because instead of the YES/NO answer to "Is the signal present?", we have six levels (NO-SURE, NO-UNCERTAIN, NO-GUESSING, YES-GUESSING, YES-UNCERTAIN, YES-SURE). These in turn give us five criterion lines (Wickens), and thus five different points of analysis when considering a user's sensitivity to a category (how far away is the user's conception of when to be sure from their conception of when to be uncertain, and so on). We can then leverage these five levels of sensitivity against where they 'should be' in nature by again applying our knowledge of category distances.

Confidence Ratings Can Help Attribute Responses to Intuited Categories

Taking confidence ratings a step further, we could even choose to give the participant the option to set their confidence in all categories. For any given trial, a participant could set their confidence to, say, 40% for one category, 25% for another category, and 35% for a third, or have any other combination of answers split up the whole of their response (user's answers couldn't total more than 100% confidence, and answers that total less than that would attribute the remaining amount of response to true uncertainty / nothing learned towards distinguishing any category from that instance). This way, we can get info about how much every category's measure of discriminability is moving from every trial, and we can compare this to how typical

each presented instance would appear in every category to tell if the participant is on track. (Though this would require drastic changes in our study as we'd pretty much have to throw response time out the window at that point).

Confidence Ratings Can Address Performance Drop in Tired Participants

Finally, confidence ratings also happen to help us in dealing with wavering performance data throughout the study. A rather obvious point we haven't addressed so far is that our study is long and people get tired. They might start either not learning or not recalling what they've learned as well as they go on. A confidence measure might help us figure out what's really going on as people lose focus and get antsy to leave, whereas how we have it now, responses become unreflective of a participant's learned state in the midst of exhaustion. By having 'guess' levels of confidence or a true center 'I don't know' option, we give the user the option to tell us that they didn't learn anything one way or the other from a trial and, in a sense, not to count it.

VI. Addressing Endurance of Learning

Another problem we've yet to mention that's truly very crucial to our goal of learning, and somewhat related to performance drop from tiredness, is that categories learned earlier in the study will often have D' values fall well below our required 2.4 criteria by the end of the study. Learners forget what they've learned when they're not studying those items.

Criterion Window

A possible way to address this issue might be in establishing a criterion window, requiring that all categories have to reach criteria in the same frame of time before a learner can truly retire any category (perhaps the window starts at the first category retired, and if every other category isn't also retired after a set number of trials, that first one is unretired, and the window begins again at the next category that was retired after it). True retirement thus occurs only when all categories are mastered at once. (If that sounds harsh, we could always make it 8/10 categories, or some number of 'most important categories').

Our ability to know and distinguish any one of these categories is inherently tied to our ability to distinguish the others—after all, what is perceiving something if not distinguishing it from other things? It's imperative we get a learner to retain their knowledge on everything learned, or at least everything important to know, by the end of the study if we want to seriously use this technology to improve education.

References

Kellman, P.J., Massey, C.M. (2013). Perceptual Learning, Cognition, and Expertise, *The Psychology of Learning and Motivation*, First Edition (2013), 117-165.

Mettler, E., Massey, C.M., & Kellman, P.J. (2011). Improving adaptive learning technology through the use of response times, *Expanding the Space of Cognitive Science: Proceedings of the 33rd annual conference of the Cognitive Science Society*, 2532–2537.

Mettler, E. & Kellman, P.J. (2014). Adaptive response-time-based category sequencing in perceptual learning, *Vision Research*, 99 (2014), 111–123.

Mettler, E., Massey, C.M., & Kellman, P.J. (2016). A Comparison of Adaptive and Fixed Schedules of Practice, *Journal of Experimental Psychology*, Vol. 145 (No. 7), 897-917.

Wickens, T.D. (2002). *Elementary Signal Detection Theory*. Oxford University Press, New York. https://doi.org/10.1093/acprof:oso/9780195092509.001.0001