Explore/exploit tradeoff strategies in a resource accumulation search task

Peter M. Todd and Ke Sang (with Robert L. Goldstone and Thomas T. Hills)

(paper at https://psyarxiv.com/zw3s8)

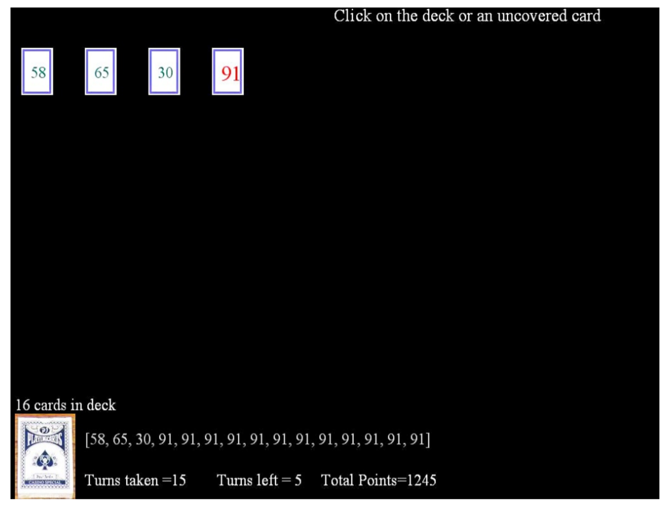
Background and Motivation

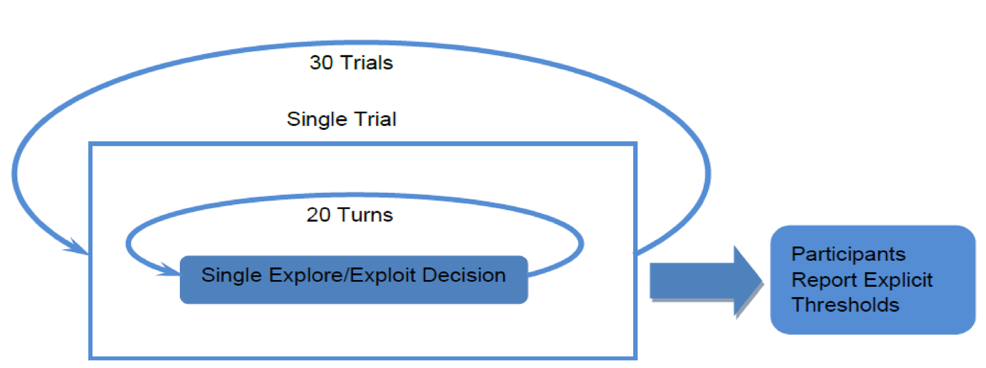
How, and how well, do people switch between exploration and exploitation to search for and accumulate resources? In a novel card selection task, participants learn to switch appropriately between exploration and exploitation and approach optimal performance. Comparing random, threshold, and sampling strategies, we find that a linear decreasing threshold rule best fits participants' behavior. Use of such rules is also supported by reaction time differences between exploration and exploitation. Decreasing threshold strategies that "front-load" exploration and switch quickly to exploitation are particularly effective in resource accumulation tasks, in contrast to optimal stopping problems like the Secretary Problem requiring longer exploration.

A Card Search Task

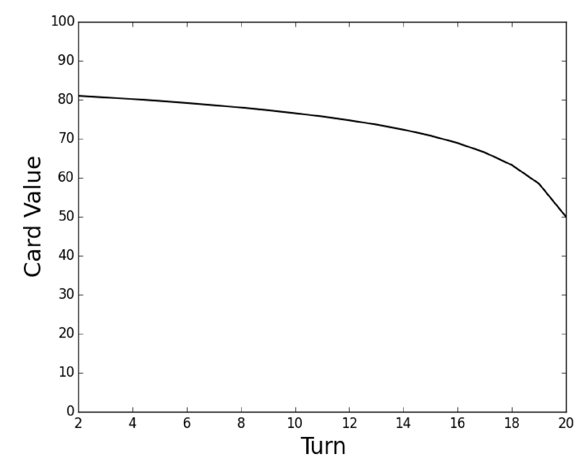
We developed a computerized card search task that allows balanced exploration and exploitation. Participants are told to accumulate as many points as possible in each of 30 trials, by turning over new cards or selecting already-found cards for 20 turns.

On the first turn of each trial participants click on the deck to start by revealing a card that is put on the "table". After that, the subject could either click on the deck again to reveal a new card (exploration), or click on one of the cards already on the table (exploitation), getting the clicked card's points. Card values range from 1 to 99, each with equal likelihood and sampling with replacement.





Card Search Task: Optimal Strategy



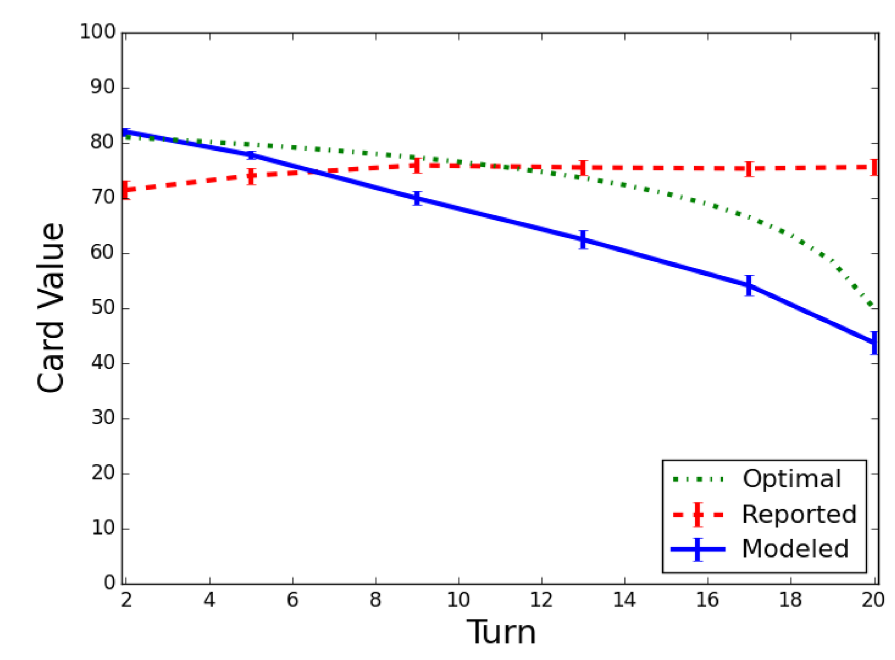
The optimal strategy requires ONLY ONE switch from exploration to exploitation, as soon as a revealed card exceeds the threshold shown at a particular turn.

Participants' Reported vs. Observed Behavior

Participants were asked what threshold they used at turns 2, 5, 9, 13, 17, and 20; their choices were also modeled with a six-threshold rule:

$$Pr_{explore}(t) = \frac{1}{1 + e^{-s[T(t) - Max]}} \qquad (1)$$

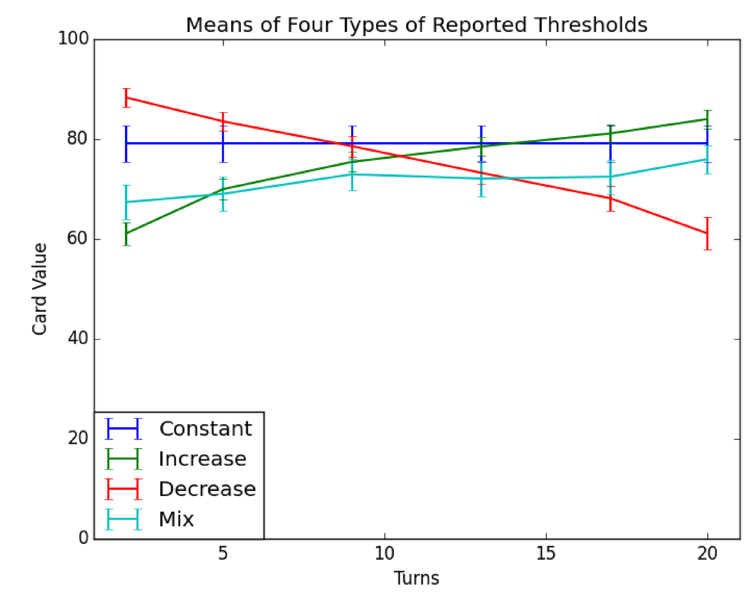
with *T(t)* being one of six thresholds corresponding to turns above, *Max* being highest card value seen so far, and *s* being a strength parameter.



Mean reported thresholds are roughly flat, but the actual choices conform to a decreasing threshold.

Individual Differences in Thresholds

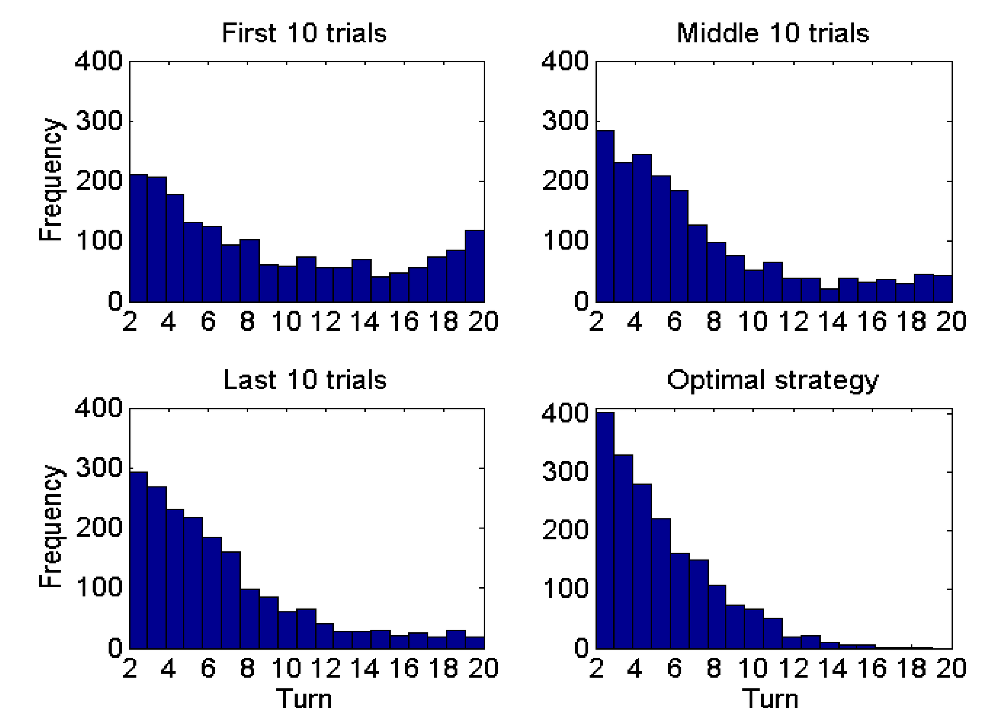
Different participants report different patterns of thresholds over turns, not just flat:



Out of N=188, 71 participants were Increasing, 47 were Decreasing, 19 were Constant (flat), and 51 were Mixed

Learning over Trials

Optimal behavior is to switch once from exploring to exploiting. Participants start off switching too late, but over 30 trials they learn to switch sooner, approaching the optimal distribution of switch points:



Participants also start off switching too often, but learn over trials to switch about once, and perform better as a result:

Trials	Number of switches	Initial turn of persistent exploitation	Performance
First 10	2.57 (1.96)	8.77 (3.63)	1492 (112.3)
Middle 10	1.54 (1.20)	6.90 (2.80)	1539 (102.6)
Last 10	1.39 (1.00)	6.39 (2.13)	1553 (91.5)
Optimal	1.0 (0.0)	5.08 (1.04)[a]	1595 (70.6)[a]

Decision Models Compared

Random baseline models:

- Epsilon-greedy (explore with probability *ε*)
- Random switch (switch to exploit at turn *k* in [2-20]

Threshold models:

- One-threshold model in eq. (1) with *T, s*
- Linear decreasing threshold, with $T(t) = b + m(t–2)$, *s*
- Two-threshold model with *T1, T2, k* (switch point), *s*
- Six-threshold model as in eq. (1)

Sampling models:

- Fixed-sample: check first *k* cards, exploit highest card seen in that sample at turn *k*+1 and rest of turns, with probability *h*
- Cutoff: check first *k* cards, set threshold at highest card seen, explore until a higher card is found and exploit it for rest of turns, with probability *h* (successful on Secretary Problem)
- Successive non-candidate count: after *j* successive non-candidates (not highest value seen) have been passed, exploit next candidate seen for rest of turns, with probability *h*

(*h* corresponds to trembling hand)

Model Comparison Results

Task performance:
Threshold models perform best, random models worst, sampling models in between (cutoff does poorly).

Fit to participant behavior:
Taking number of parameters into account via BIC, the **linear decreasing threshold** fit behavior best along with other threshold rules; the **cutoff** model also did well.

In too much detail:

	Best performing model			Best fitting model to data			
Strategy	Score per trial	Best parameter values	Switch turn (and % switching)	Best fit parameter values	Best fit error parameter	Number of parameters	BIC
Participant performance	1528.0	NA	7.64 (94.9%)	NA	NA	NA	NA
Optimal	1601.8	NA	5.51 (100%)	NA	$s = 0.12$	1	431.5
Epsilon-greedy	1312.0	$\varepsilon = 0.34$	NA	$\varepsilon = 0.21$	NA	1	596.5
Random switch	1318.9	NA	7.62 (95.3%)	NA	$s = 0.21$	1	454.8
One-threshold	1599.0	$T = 79$	5.50 (99.0%)	$T = 68.3$	$s = 0.132$	2	378.3
Linear decreasing threshold	**1601.7**	$m = -0.58$ $b = 81$	5.47 (99.9%)	$m = -1.78$ $b = 80.65$	$s = 0.12$	3	**326.4**
Two-threshold	1601.1	$T_1 = 80$ $T_2 = 75$ $k = 8$	5.44 (99.7%)	$T_1 = 77.1$ $T_2 = 57.7$ $k = 7$	$s = 0.126$	4	335.7
Six-threshold	1601.7	$T_1 = 82$ $T_2 = 81$ $T_3 = 79$ $T_4 = 76$ $T_5 = 71$ $T_6 = 58$	5.55 (100%)	$T_1 = 82.0$ $T_2 = 77.7$ $T_3 = 69.9$ $T_4 = 62.5$ $T_5 = 54.1$ $T_6 = 44.1$	$s = 0.13$	7	346.8
Fixed sample	1495.7	$k = 6$	7.0 (100%)	$k = 4$	$h = 0.42$	2	445.6
Cutoff	1391.6	$k = 2$	6.60 (90.4%)	$k = 2$	$h = 0.095$	2	389.5
Successive non-candidate count	1469.9	$j = 3$	7.29 (99.99%)	$j = 1$	$h = 0.46$	2	592.7

The majority of individual participants were fit best by the linear decreasing threshold model, but many were best fit instead by the cutoff model, indicating they may have thought (wrongly) of the task as a Secretary Problem:

Strategy	Number of participants	Participants' mean score
Linear decreasing threshold	117	1545
Cutoff	38	1486
Two-threshold	24	1531
One-threshold	9	1532
Random switch	2	1415
Fixed sample	1	1306
Six-threshold	0	NA
Epsilon-greedy	0	NA
Successive non-candidate count	0	NA

Conclusions

- **In resource-accumulating search, it is best to pick an option to exploit early on, so start choosy and quickly drop in choosiness, as decreasing thresholds do**
- **This differs from optimal stopping search with one-time payoffs, calling for longer exploration**
- **Participants are mostly sensitive to the difference, using decreasing threshold rules in the card search task, though they cannot explicitly indicate their search rule**
- **Participants learn to do better over trials, switching to exploitation sooner and less frequently**
- **Response times (not shown) indicate faster exploit than explore decisions, also supporting decreasing thresholds.**

Acknowledgments: Thanks to Ross Branscombe, Jerome R. Busemeyer, and Woo-Young Ahn for help with this research, and to the National Science Foundation REESE grant 0910218 and the John Templeton Foundation grant "What drives human cognitive evolution" for support.