Feedback Promotes Learning and Knowledge of the Distribution of Values Hinders Exploration in an Optimal Stopping Task

Erin H. Bugbee (ebugbee@cmu.edu)
Department of Social and Decision Sciences, Carnegie Mellon University
Pittsburgh, PA 15213 USA

Cleotilde Gonzalez (coty@cmu.edu)
Department of Social and Decision Sciences, Carnegie Mellon University
Pittsburgh, PA 15213 USA

Abstract

People frequently encounter the challenge of deciding when to stop exploring options to optimize outcomes, such as when selecting an apartment in a fluctuating housing market or booking a dinner reservation on New Year's Eve. Despite experiencing these decisions on multiple occasions, people often struggle to stop searching optimally. This research investigates human learning abilities in optimal stopping tasks, focusing on feedback and knowledge of option value distributions. Through an experimental sequential choice task, we demonstrate that experience improves performance, with feedback significantly influencing learning. We also find that awareness of the value distribution reduces the duration of the search. A cognitive model accurately predicts these effects, shedding light on human learning processes.

Keywords: sequential decision making; optimal stopping; learning from experience; exploration; feedback

Introduction

Sequential stopping tasks and, in particular, optimal stopping tasks are crucial to understanding decision making in contexts where people must decide when to stop searching for better options. Previous work indicates that people often stop earlier than optimal (e.g. Campbell & Lee, 2006; Guan, Stokes, Vandekerckhove, & Lee, 2020; Baumann, Singmann, Gershman, & von Helversen, 2020; Lee & Courey, 2021). This work has also emphasized that participants may not be able to learn from experience to stop more optimally. As a result, there is very little research on learning to stop from experience and the contextual factors that can influence learning and stopping behavior.

Recent work has provided initial evidence that people can learn to stop closer to optimal from experience (Goldstein, McAfee, Suri, & Wright, 2020), and that people adapt to manipulations in context (Baumann, Schlegelmilch, & von Helversen, 2022). However, these previous experiments had many differences, including the context of the decision, making it difficult to generalize the factors that influence stopping behavior and determine when people learn and what factors influence learning. Many open questions remain about how people decide when to stop, how they learn when to stop, and what influences learning and stopping behavior.

Our research program aims to understand when and how people learn when to stop searching in sequential choice tasks. We do so by creating a new optimal stopping task that we use to directly manipulate factors and observe their effect on learning and stopping behavior in humans independent of context. We also create a cognitive model that replicates human learning behavior.

We begin by investigating two factors that have varied between optimal stopping experiments and for which we expect an influence on learning. One difference between the previous experiments is the feedback provided, with some experiments providing no feedback (e.g. Lee & Courey, 2021), some providing outcome feedback with only the correctness of their decisions (e.g. Guan et al., 2020), and others providing varying degrees of detailed feedback (e.g. Baumann et al., 2022; Goldstein et al., 2020). Campbell and Lee (2006) manipulated feedback, but found no evidence of learning, although the distribution of values that people would experience was always known to participants.

In fact, experiments have also varied in the degree of knowledge people have about the distribution of option values. Some experiments have provided participants with a learning phase (e.g. Baumann et al., 2022) or have informed them of the distribution (e.g. Campbell & Lee, 2006), while Goldstein et al. (2020) required participants to learn the distribution purely from experience.

We investigate how people can learn when to stop exploring options and make a selection based on the form of feedback they receive and their knowledge of the distribution of option values. We consider the following: How do feedback and knowledge of the distribution of options values affect learning in sequential search? How do people deviate from optimal based on these factors, and can this be modeled with a cognitive model of decisions from experience?

This study builds on previous research by integrating different factors across tasks into one task and a single experiment, while varying feedback and knowledge of the distribution of values to determine their influence on the decision of when to stop. We hypothesized that people would learn from experience to improve their accuracy over time and that people would achieve highest accuracy when receiving detailed feedback, followed by outcome feedback, then no feedback. We also hypothesized that people would learn more when they did not know the distribution of option values, although knowing may lead to better performance. In line with previous work on Instance-Based Learning (IBL) models of sequential decisions (Bugbee & Gonzalez, 2022; Bugbee, McDonald, & Gonzalez, 2022), we predicted that an IBL model would accurately emulate human stopping behavior.

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

Experiment

This experiment investigates how feedback and knowledge of the distribution of option values affect decision-making processes in an optimal stopping task. The task required participants to select a box or pass it with the caveat that once passed that box could not be selected later. If the last box in a sequence was reached, the participants were forced to select it. We manipulated their awareness of the value distribution (Known or Unknown) and varied the feedback given to participants (No Feedback, Outcome, or Detailed). The box values were sampled without replacement from a truncated normal distribution $\mathcal{N}(50, 20)$ bounded from $[0, 100]$.

Methods

Participants were recruited through Amazon Mechanical Turk. They were paid \$3.00 for completing the task and \$0.02 for each correct problem, with the opportunity to earn up to \$1.00 in bonus payment. Participants were removed from the analysis if they attempted any part of the task multiple times or if we did not obtain complete data from them, leaving a total of 256 participants for analysis. Participants were 39.5% women, 58.2% men, and 2.3% non-binary or preferred not to answer. The median age was 38 years ($SD = 10$).

The experiment had six conditions in a 2 (Knowledge of Distribution: Known or Unknown) x 3 (Feedback: No Feedback, Outcome, Detailed) between-subjects design. Participants were randomly assigned to a condition with 44 participants in Known, No Feedback; 46 in Known, Outcome; 41 in Known, Detailed; 42 in Unknown, No Feedback; 40 in Unknown, Outcome; and 43 in Unknown, Detailed.

Procedure and Optimal Stopping Task Participants began by completing a Qualtrics survey where they were given instructions about the task and learned the distribution if in the Known condition. Then, they were redirected to a novel optimal stopping task. In the task, each participant completed 50 problems. For each problem, participants were shown the value of one box at a time in a sequence, as shown in Figure 1. Each problem consisted of 10 boxes. Participants had to make a choice to "Pass" and move on to the next box, or "Select" the current box and terminate search in that problem if they believed it was the maximum value in that sequence. Search ended when either a selection was made or when the last box (Box 10) in the sequence was reached and participants were forced to select it. After completing all 50 problems, they were redirected back to Qualtrics to answer some demographic and task-related questions.

Knowledge of Distribution of Option Values In the Known condition, participants were informed of the distribution of box values before starting the task. The distribution was shown graphically alongside a verbal explanation of how to understand the graph. The participants then answered a series of questions about the distribution's minimum, maximum, median, most likely range, and least likely ranges to ensure that they understood the distribution of values they would experience. They received the correct answers with explanations after answering each question. In the Unknown condition, participants were not informed of the distribution and started the task without this knowledge.

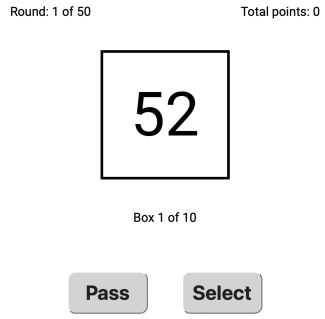
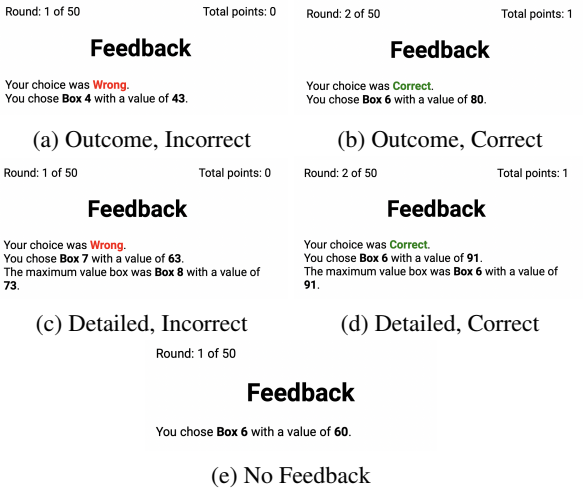


Figure 1: Participants were shown a box and chose to "Pass" and continue to the next box or "Select." Upon selection, they received condition-dependent feedback (see Figure 2).

Feedback In the No Feedback condition, participants did not receive any feedback on their selections, only the box they chose and its value. In the Outcome Feedback condition, participants were informed about the outcome of their selection, which was whether they correctly selected the maximum. In the Detailed Feedback condition, participants received detailed information, including the correctness of their selection, as well as which box was the maximum and its value. In the Outcome and Detailed Feedback conditions, the total points earned was displayed in the upper right corner. Feedback as shown to the participants is shown in Figure 2.



(a) Outcome, Incorrect

(b) Outcome, Correct

(c) Detailed, Incorrect

(d) Detailed, Correct

(e) No Feedback

Figure 2: Feedback by condition and correctness of selection.

Models

Instance-Based Learning Model

We built an Instance-Based Learning (IBL) model that makes sequential decisions according to a theory of decisions

from experience, Instance-Based Learning Theory (IBLT) (Gonzalez, Lerch, & Lebiere, 2003). These models have been shown to be predictive in sequential decision tasks (Bugbee & Gonzalez, 2022; Bugbee et al., 2022). We briefly summarize the theory (for the mathematical algorithm, see Nguyen, Phan, & Gonzalez, 2022; Gonzalez, 2023).

IBLT proposes that learning occurs through the accumulation of memory units called instances. Each instance represents a potential decision or a decision made, and each instance has an activation value that represents the ease of retrieval of that information from memory, according to the Activation equation from ACT-R (Anderson & Lebiere, 2014). Past instances are retrieved according to their similarity to the current situation, their frequency of occurrence, and their recency. A blended value (BV) for each of the two alternatives (Pass and Select) is a form of expected utility, calculated as the sum of the product of the probability of retrieval of each instance and its utility. For each decision, the IBL agent chooses the alternative with the highest BV. When the agent receives feedback, this is stored as the utility for that instance.

The instance structure that we designed for this model is shown in Table 1. The state consists of the value of the box and the number of boxes remaining in the sequence after the current box; the action is to Select or Pass; and the utility is binary. For the No Feedback condition, the utility corresponds to whether the box chosen, if after the first box, is the best so far and thus has the possibility of being the maximum. If the box is the first box or is not the maximum, the utility is 0 because there is no information that indicates that it could be the maximum. If the box is not first and is the best so far, the utility is 1. For both feedback conditions, the utility is based on the correctness of the selection, with a 1 if the maximum box is correctly selected and 0 otherwise. For the Detailed Feedback condition, there are additional instances for selecting the correct box as provided in the feedback and for passing all encountered boxes that were not the maximum. This is because the feedback provides information on both the value and position of the box that should have been selected and those seen that should not have. These additional instances have a utility of 1, since they indicate what actions should have been taken.

Table 1: Instance Structure

State		Action	Utility
Value	Boxes Remaining	{Select, Pass}	{0, 1}

We use linear similarity to compare each of the attributes of the current state and the state of past instances, and the decay and noise parameters are set to the ACT-R default values of $d = 0.5$ and $\sigma = 0.25$ respectively. The model also uses credit assignment (Nguyen et al., 2022), assigning an expected value, which is the blended value for passing at that position, to all pass actions until a selection is made, and then all decisions are updated to the outcome obtained.

Before beginning the task, we provide agents with prior knowledge in the form of prepopulated instances. We populate all agents with instances giving a utility of 2 to selecting 100 in the first box and a 0 to passing it, and a 0 to selecting 0 in the first box and 2 for passing it. This is to represent knowledge people may have, such as knowing that higher values are better than lower ones. The utilities are higher than what can possibly be obtained to encourage exploration of both actions.

Additionally, agents in the Known condition learn the distribution before starting the task. These correspond to the information that humans have learned about the distribution. We populate the agents in the Known condition with instances indicating that the values of 100, 90, and 80 can lead to a utility of 1 if selected and 0 if passed, and that the values of 0, 10, 20, and 30 can give a utility of 0 if selected and 1 if passed. We add these instances for every position, so that there are instances for each value of the "boxes remaining" attribute.

We simulated an *IBL Agent* for each human participant. That is, we have a corresponding agent that completes the same problems as that participant. The number of agents is the same as the number of participants in each condition.

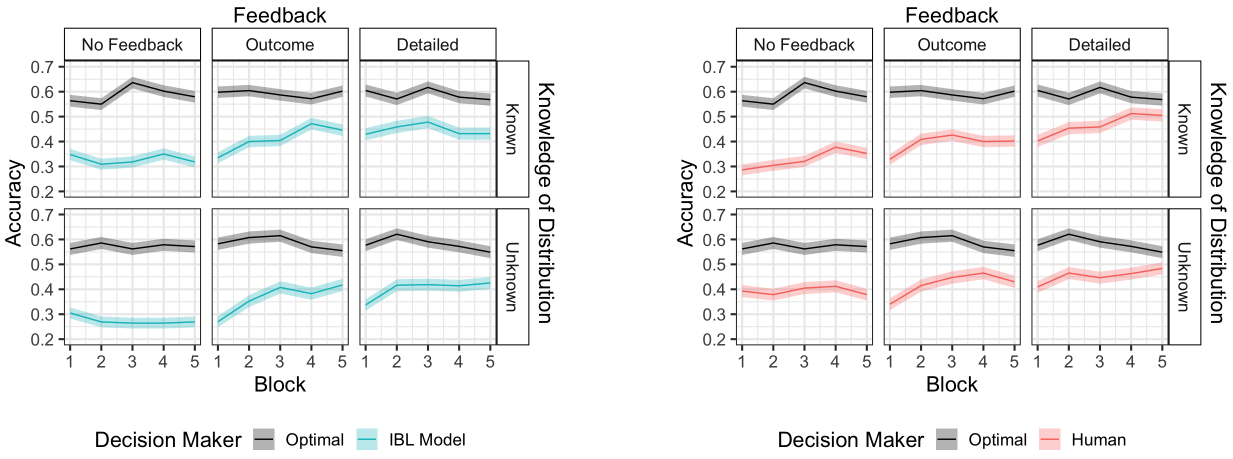
Optimal Model

Given a particular sequence length and distribution, there is an optimal solution that involves following optimal thresholds that depend on the position. These thresholds can be calculated following the process proposed by Gilbert and Mosteller (1966). We calculate the optimal thresholds using the percentiles from Table 7, Column 2 of Gilbert and Mosteller (1966), in alignment with Table 1, Column 2 from Goldstein et al. (2020). We simulated an *Optimal Agent* that follows the optimal thresholds for each human participant: that is, the agent compares the value of the current box to the optimal threshold for that position in the sequence, and if the value exceeds the threshold and is the best so far, it selects it, otherwise it passes. Optimal agents do not learn; they deterministically decide when to stop based on the optimal thresholds. We simulated an optimal agent for each human participant which encountered the same sequences of box values in the same order as the human.

Predictions of the IBL Model

For each human participant, there is an IBL agent and an optimal agent completing the same problems. The results of simulations with IBL agents serve to predict human behavior, and those with optimal agents provide a benchmark for what humans can aspire to if they learn the distribution of values and optimal thresholds. The optimal strategy is the same for all conditions because the distribution and sequence length are the same, resulting in equivalent optimal thresholds; however, there are minor differences in results due to the randomness of sampling values from the distribution. The results are aggregated into 5 blocks of 10 problems each.

We investigate differences in performance by condition. Performance is assessed by accuracy, the proportion of problems in which the maximum box was selected, and stopping errors, the proportion of problems in which the selected box was before or after the maximum.



(a) IBL Agents

(b) Human Participants

Figure 3: Accuracy for IBL agents (3a) and human participants (3b) alongside optimal agents, over blocks of 10 problems by feedback and knowledge of the distribution. Error bars represent standard errors for the mean.

Accuracy

Figure 3a shows the accuracy for the optimal and IBL agents. Optimal agents achieve an average accuracy of 0.58 ($SE = 0.004$). This indicates that by following the optimal strategy, one can expect to be correct on about 60% of the problems.

The prediction based on IBL agents is that the accuracy will be highest when receiving detailed feedback, followed by outcome feedback, and then no feedback. Humans are also expected to improve their accuracy with experience when receiving feedback. Lower accuracy is expected initially when the distribution is unknown, though this effect is minor.

Stopping Errors

Figure 4a displays the stopping errors for optimal and IBL agents, which are made by stopping before or after the maximum value box. Following the optimal strategy can lead to errors, and optimal agents have slightly more early stopping errors than late stopping errors.

IBL agents stop before the maximum substantially more often than is optimal. With experience, these early stopping errors decrease when receiving feedback, approaching the optimal proportion of early stopping errors. Errors made by stopping after the maximum are closer to optimal in all conditions, and are slightly lower than is optimal in most conditions. IBL agents predict that, under conditions that involve feedback, humans will approach optimal errors with experience by reducing early stopping errors over blocks.

Search Length

The search length is the total number of boxes searched to make a selection. For example, if the third box is selected, then the search length is 3.

Figure 5a shows the search length for optimal and IBL agents. The optimal average search length is consistently slightly below 6 across all conditions and problems. The IBL agents search substantially less than is optimal. The search length is predicted to be stable and the shortest when receiving no feedback. When receiving feedback, the agents suggest that humans will start with a short search length but will increase its duration over blocks, approaching the optimal search length. Feedback is expected to encourage exploration, helping agents learn to search longer, while no feedback stifles exploration.

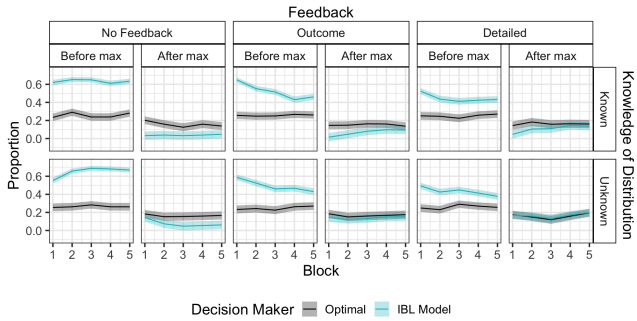
We also observe a longer search length when the distribution is unknown relative to known. Thus, having knowledge of the distribution may hinder exploration, resulting in search lengths that are further from optimal.

Human Experimental Results

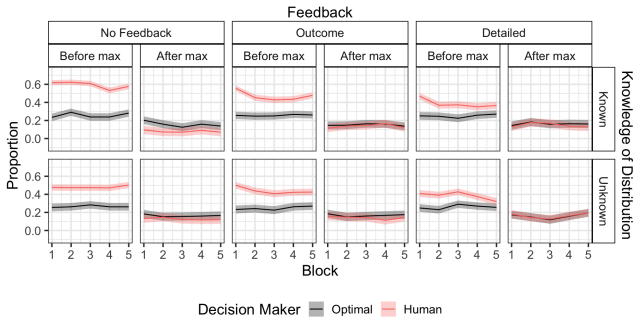
Accuracy

Figure 3b shows the average accuracy over blocks across all participants. As predicted by the IBL model, we observe that participants improved their accuracy over time when they received feedback, approaching optimal accuracy. We do not observe a significant effect of the knowledge of the distribution, also as predicted by the IBL agents.

To determine if people learn to improve their accuracy with experience, we ran a repeated-measures ANOVA predicting accuracy with feedback and knowledge of the distribution as between-subjects factors and block as within-subjects. As shown in Table 2, the ANOVA indicates that feedback and block had a statistically significant effect on accuracy. When comparing accuracy by block, we observe that accuracy increases over blocks, implying that there was substantial learning from the beginning to the end of the task. Also, participants obtained the highest accuracy when receiving detailed ($M = 0.46, SE = 0.008$) relative to outcome feedback ($M = 0.41, SE = 0.007$) and outcome relative to no feedback ($M = 0.36, SE = 0.007$). In addition, they had slightly



(a) IBL Agents



(b) Human Participants

Figure 4: Stopping errors calculated as the proportion of problems in which the decision maker stopped before or after the maximum for IBL agents (4a) and human participants (4b) alongside optimal agents, over blocks of 10 problems by feedback and knowledge of the distribution. Error bars represent standard errors for the mean.

higher accuracy when the distribution was unknown ($M = 0.42, SE = 0.006$) relative to known ($M = 0.39, SE = 0.006$), though this effect is not significant.

Table 2: Repeated-measures ANOVA predicting accuracy with feedback and knowledge between-subjects and block within-subjects.

Source	Df	F value	$Pr(>F)$
Error: Participant			
Feedback	2	7.81	< 0.001 ***
Knowledge	1	1.66	0.199
Feedback:Knowledge	2	1.19	0.306
Residuals	250		
Error: Participant:Block			
Block	4	9.66	< 0.001 ***
Feedback:Block	8	1.34	0.220
Knowledge:Block	4	0.41	0.804
Feedback:Knowledge:Block	8	0.82	0.584
Residuals	1000		

Stopping Errors

Figure 4b displays stopping errors over time for human participants. As predicted by IBL agents, human participants make more early stopping errors than is optimal. In contrast, the late errors are lower and close to those committed by the optimal agents. With outcome and detailed feedback, human participants decrease their early stopping errors over blocks, whereas no such decrease is observed in the no feedback condition, as predicted by IBL agents.

Search Length

Figure 5b shows the average search length over blocks of problems and Table 3 shows the results of a repeated-measures ANOVA predicting search length with feedback and knowledge between-subjects and block within-subjects.

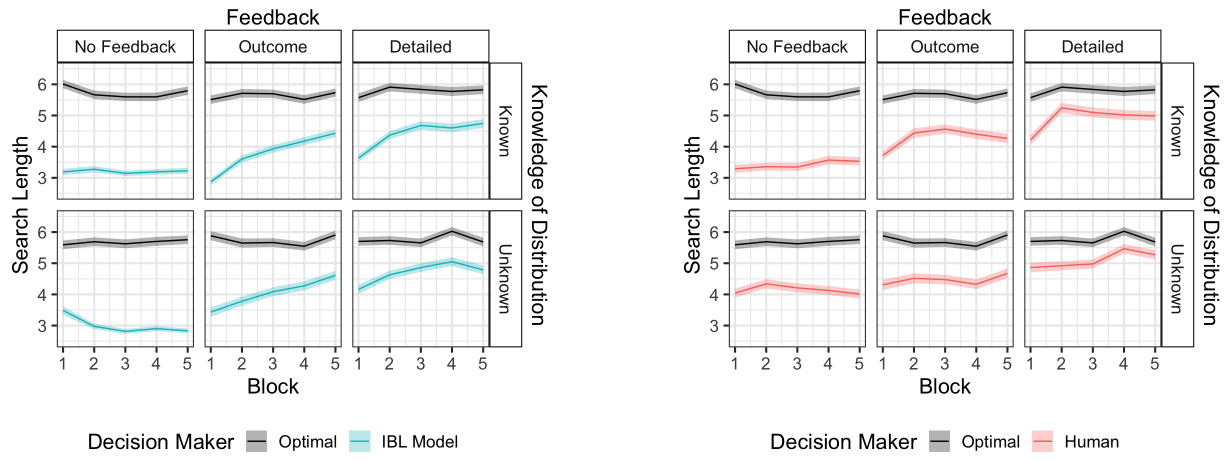
We observe that feedback affects the length of the search. People search more when receiving detailed ($M = 5.01, SE = 0.049$) relative to outcome feedback ($M = 4.36, SE = 0.049$), and when receiving outcome relative to no feedback ($M = 3.77, SE = 0.044$). Knowledge of the distribution significantly affects the length of the search, with people searching more when the distribution of values is unknown ($M = 4.57, SE = 0.039$) relative to known ($M = 4.19, SE = 0.039$).

We also see that the block significantly affects search length, indicating that people are learning when to stop. Search duration increases with experience. When receiving detailed feedback, the increase from block 1 ($M = 4.55, SE = 0.111$) to block 5 is substantial ($M = 5.13, SE = 0.108$), as is the increase from block 1 ($M = 3.99, SE = 0.106$) to block 5 ($M = 4.46, SE = 0.109$) when receiving outcome feedback. With no feedback, the increase from block 1 ($M = 3.66, SE = 0.098$) to block 5 ($M = 3.77, SD = 0.099$) is less notable.

Finally, we find a significant three-way interaction of feedback, knowledge, and block. The impact of knowledge and block depends on feedback; when receiving detailed feedback, block and the interaction of knowledge and block are significant; for outcome feedback, block is significant; and for no feedback, knowledge is significant.

Table 3: Repeated-measures ANOVA predicting search length with feedback and knowledge between-subjects and block within-subjects.

Source	Df	F value	$Pr(>F)$
Error: Participant			
Feedback	2	16.48	< 0.001 ***
Knowledge	1	4.39	0.037 *
Feedback:Knowledge	2	1.07	0.346
Residuals	250		
Error: Participant:Block			
Block	4	7.28	< 0.001 ***
Feedback:Block	8	1.17	0.312
Knowledge:Block	4	1.86	0.116
Feedback:Knowledge:Block	8	1.96	0.049 *
Residuals	1000		



(a) IBL Agents

(b) Human Participants

Figure 5: Search length for IBL agents (5a) and human participants (5b) alongside optimal agents, over blocks of 10 problems by feedback and knowledge of the distribution. Error bars represent standard errors for the mean.

Learning to Stop Optimally: Lessons from an IBL Model

We have provided new evidence for the ability of humans to learn to improve stopping accuracy with experience. Furthermore, given that the theoretical predictions from the IBL model are reflected in empirical data, the IBL model can provide insights into how and what people learn.

The IBL model follows a computational algorithm that represents the process we may apply in our minds when using our past experiences to make these stopping decisions. More concretely, the representations we have chosen in the model may help explain the differences in human behavior according to the condition.

As predicted by IBL agents, humans achieve the highest accuracy when receiving detailed feedback. Receiving information on the correctness of their selection as well as the true correct decision and that box's position and value suggest that humans may be learning not only how their chosen boxes and their corresponding position and value translate into a reward, but also what values for each position tend to be correct. For the outcome feedback condition, people receive only correctness information about their own selections and thus learn how the chosen boxes relate to outcomes but not what was correct if they chose incorrectly. For the no feedback condition, utility is encoded as whether the model chose the best value so far after the first box since the first box is always the best. Human participants likely consider how their chosen boxes relate to the previous boxes in the sequence.

The model also generally predicts the finding that people search more when the distribution is unknown. Prior knowledge of the distribution is encoded in the model's prepopulated instances, and so this appears to be affecting search decisions by hindering exploration.

When the distribution is unknown and there is no feedback, the search length is predicted to be shortest by the IBL model but humans searched longer than expected. This suggests that people may be using some other form of prior knowledge that we do not include in our model. Future research should help us understand how people explore when they do not receive feedback and do not know anything about the environment.

Conclusion

The decision of when to make a selection while observing a sequence of alternatives is faced in people's daily lives. These decisions are difficult and require the prediction of the values of future options along with the impossibility of returning to foregone options. With the difficulty and prevalence of these decisions, it would be beneficial for people to learn to make better stopping decisions with experience.

This research provides novel and unique evidence that people can learn from experience to improve and achieve near-optimal stopping in a sequential decision task. We demonstrate the role that feedback plays on this learning and how, with feedback, people significantly improve their accuracy with experience.

We also observe that having knowledge of the distribution of option values hinders exploration and may negatively impact accuracy slightly. When people do not have this knowledge, they need to explore to discover the possible option values, and since people tend to make more early stopping errors than late ones, this brings them closer to the optimal stopping point and increases their accuracy.

Importantly, the behavioral phenomena discovered in this investigation are nicely captured by the "out-of-the-box predictions" of a theoretical model of decision making based on experience. This IBL model provides insight into how and what people learn when deciding when to stop, and future work will investigate this connection further.

Acknowledgments

An Open Science Framework project is available at https://osf.io/bqxhz with data, code, and analysis files. This work was supported by the NSF AI Institute for Societal Decision Making (AI-SDM) under grant number IIS 2229881 and by the Air Force Research Laboratory under grant number 15742239.

References

Anderson, J. R., & Lebiere, C. J. (2014). *The atomic components of thought*. Psychology Press.

Baumann, C., Schlegelmilch, R., & von Helversen, B. (2022, October). Adaptive behavior in optimal sequential search. *Journal of Experimental Psychology: General*. doi: 10.1037/xge0001287

Baumann, C., Singmann, H., Gershman, S. J., & von Helversen, B. (2020, June). A linear threshold model for optimal stopping behavior. *Proceedings of the National Academy of Sciences*, *117*(23), 12750–12755. doi: 10.1073/pnas.2002312117

Bugbee, E. H., & Gonzalez, C. (2022). Making predictions without data: How an instance-based learning model predicts sequential decisions in the balloon analog risk task. In *Proceedings of the 44th annual meeting of the cognitive science society* (p. 3167-3174). Cognitive Science Society.

Bugbee, E. H., McDonald, C., & Gonzalez, C. (2022). Leveraging cognitive models for the wisdom of crowds in sequential decision tasks. In *Paper presented at Virtual MathPsych/ICCM 2022* (p. 7). Retrieved from mathpsych.org/presentation/751

Campbell, J., & Lee, M. D. (2006). The effect of feedback and financial reward on human performance solving 'secretary' problems. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 28).

Gilbert, J. P., & Mosteller, F. (1966, March). Recognizing the Maximum of a Sequence. *Journal of the American Statistical Association*, *61*(313), 35–73. doi: 10.1080/01621459.1966.10502008

Goldstein, D. G., McAfee, R. P., Suri, S., & Wright, J. R. (2020, March). Learning when to stop searching. *Management Science*, *66*(3), 1375–1394. doi: 10.1287/mnsc.2018.3245

Gonzalez, C. (2023, October). Building human-like artificial agents: A general cognitive algorithm for emulating human decision-making in dynamic environments. *Perspectives on Psychological Science*, 17456916231196766. doi: 10.1177/17456916231196766

Gonzalez, C., Lerch, J. F., & Lebiere, C. (2003, July). Instance-based learning in dynamic decision making. *Cognitive Science*, *27*(4), 591–635. doi: 10.1207/s15516709cog2704$_2$

Guan, M., Stokes, R., Vandekerckhove, J., & Lee, M. D. (2020, September). A cognitive modeling analysis of risk in sequential choice tasks. *Judgment and Decision Making*, *15*(5), 823–850. doi: 10.31234/osf.io/evzp9

Lee, M. D., & Courey, K. A. (2021, March). Modeling optimal stopping in changing environments: A case study in mate selection. *Computational Brain & Behavior*, *4*(1), 1–17. doi: 10.1007/s42113-020-00085-9

Nguyen, T. N., Phan, D. N., & Gonzalez, C. (2022, June). SpeedyIBL: A comprehensive, precise, and fast implementation of instance-based learning theory. *Behavior Research Methods*. doi: 10.3758/s13428-022-01848-x