EXTENDING EVIDENCE ACCRUAL MODELS OF TWO-ALTERNATIVE FORCED-CHOICE DECISION MAKING TO THE n-ALTERNATIVE CASE

Steven R. Carroll, William M. Petrusic
Carleton University
Ottawa, Ontario, Canada
srcarrol@connect.carleton.ca

Abstract

Stimuli in a series of four experiments were 20x20 matrices, each element of which was a square coloured red, green, blue, or black. Systematic manipulation of the frequencies with which each colour appeared allowed for the generation of two, three, and four coloured stimuli. By asking participants to evaluate the stimuli and decide which colour was represented either 'most' or 'least', response time, accuracy, mean confidence, and mean time to render confidence data was generated for a series of n-alternative forced-choice (nAFC) decisions where 2≤n≤4. A new nAFC model is discussed in light of these data.

Many modellers of two-alternative forced choice (2AFC) sensory-based decision-making assume judgements follow from a series of discrete evidence accrual events (for example, Vickers, 1979; Petrusic, 1992; Van Zandt, 2000). Under these models, a decision is made only after a criterion amount of evidence has been collected either in support of one possible alternative choice or the other.

Few researchers have attempted to extend these models beyond the limited 2AFC case and into the realm of nAFC decision-making. Vickers (1979; see also Vickers and Lee, 1998) is the notable exception and, as such, his evidence accrual model of nAFC sensory-based decision-making shall be used as the exemplar for this entire class of models in the discussion which follows.

According to Vickers, nAFC decision-making is entirely comparable to 2AFC decision-making. Sampled evidence is evaluated with regards to 'n' different permissible response categories and the appropriate evidence counter increments. When any of these counters reaches a criterion evidence level a decision is made.

Vickers postulates that a decision-maker uses mean confidence in the integrity of the decision-making process to adjust these criterion levels. Low confidence will invoke stricter criteria and, consequently, a comparable increase in decisional response time (RT). Comparably, overly high confidence will cause criterion levels, and RT, to decrease. It follows from this argument that, since decision-makers are attempting to achieve an ideal level of confidence and are able to adjust criteria to meet this end, overall mean confidence should eventually stabilise at this ideal level regardless of the size of 'n' in a series of nAFC decisions.

Since no one has yet gathered mean RT and mean confidence data within the confines of an experiment which systematically manipulates 'n', it is unclear whether Vickers predictions regarding a stable confidence level would materialize. One purpose of the present study is to test this assumption. Further, the present study will test whether Vickers' predictions are consistent with observations made by Hick (1952), who showed that the difference between mean decisional response times increase along with decisional complexity, but in a non-linear fashion. Specifically, Hick has demonstrated how the differences between mean RTs for a series of nAFC decisions and a series of n+1AFC decisions tends to decrease as 'n' increases.

General Method

Participants

96 undergraduate psychology students from Carleton University participated in these studies in return for course credit. Twenty-four participants were assigned to each of four experiments. All participants were tested a priori for color blindness.

Stimuli

Stimuli in all four experiments were 20 x 20 arrays of squares. Each square was coloured red, green, blue, or black. Stimuli were either two coloured, three coloured, or four coloured. In order that any effect of increasing the number of possible colour response alternatives (a variable hereafter referred to as “decisional complexity”) not be confounded with effects of increases in decisional difficulty, the ratio between groups of coloured squares was held as nearly as possible to 1:1.08 for all stimuli in Experiment 1. In the remaining experiments, the between colour ratios were manipulated in order to create three levels of decisional difficulty within each level of decisional complexity. "Easy" decisions involved ratios held as nearly as possible to 1:1.16, while "hard" decisions involved a ratio of 1:1.04. "Medium" decisions used the same colour difference ratio as was used in Experiment 1.

Procedure

Participants were seated before a computer monitor, within easy reach of a response panel. The labels “BLACK”, “RED”, “GREEN”, and “BLUE” written in appropriately coloured ink beneath four keys arranged along the bottom of the response panel, and the labels “X/GUESS/LOW/MOD LOW/MOD HIGH/HIGH/CERTAIN” written in black ink beneath the seven keys aligned along the top of the response panel.

Each trial of each experiment began with the presentation of an instruction: "MOST" or "LEAST" for all four experiments, and the instructions "2^{nd} MOST" and "2^{nd} LEAST" were used for half of the trials in the fourth experiment. The instruction was centred along the bottom of the computer monitor. 1 second after instruction presentation one of the coloured stimuli appeared centered above the instruction. The participants' task was to select the colour which corresponded to the presented instruction. For example, if the instruction was "MOST" participants would choose the colour they thought appeared most often in the display.

Following their decision, the stimulus and instruction were removed from the display and participants were prompted to express their confidence in having made a correct decision by selecting one of the seven confidence response keys.

In Experiment 1 participants made decisions in three blocks of trials: One where stimuli were two coloured, a second where stimuli were three coloured, and a fourth where stimuli were four coloured. 48 stimuli were created for each block resulting in 144 stimuli total. All decisions made in this experiment were of "medium" difficulty (as defined above). Block order was randomized between participants.

Experiment 2 replicated Experiment 1, but added two additional levels of decisional difficulty ("easy" and "hard" as defined above). 48 stimuli were created for each level of difficulty for each block of trials, resulting in 432 stimuli total.

Experiment 3 replicated Experiment 2, but all subjects were asked to participate in a second session wherein they were not asked to render confidence judgements following each decision, but which was otherwise identical to the first session (in total, 864 decisions per participant). Session type order was randomized between participants, as was block order.

Experiment 4 was comparable to Experiment 2, but all subjects were asked to participate in a second session wherein the instructions were changed from "MOST/LEAST" to "2nd MOST/2nd LEAST" but which was otherwise identical to the first session (in total, 864 decisions per participant). Again, session type order was randomized between participants, as was block order.

Results

For all experiments, trials where RTs were less than 200 ms in length were censored, as were trials where RTs more than 3 standard deviations above each participant's mean RT within each block. This accounted for 1.88%, 2.71%, 2.09%, and 2.15% of the data in experiments 1-4 respectively. In the following ANOVAs Huynh-Feldt epsilon adjustment degrees of freedom are used, but the degrees of freedom reported are those defined by the design. An $\alpha_{pc} = .05$ is used as an index of significance in all reported analyses. In each case block order serves as a between participant independent variable (IV), while decisional difficulty, condition, and nAFC served as within participant IVs. Relevant dependant variables (DVs) were mean decisional response time (RT), arc-sin transformed mean probability of a correct response (accuracy), mean time to render post-decisional confidence ratings (CT), and mean confidence ratings.

Decisional Response Time Analyses

Table 1. RT main effect analyses for each of the relevant experiments

Factor		Experiment number and Result		Experiment number and Result
nAFC	1	$F(2,36) = 21.63$, $p \leq .001$, $\mu_p^2 = .55$	2	$F(2,36) = 13.12$, $p \leq .001$, $\mu_p^2 = .42$
	3	$F(2,24) = 25.04$, $p \leq .001$, $\mu_p^2 = .68$	4	$F(2,24) = 23.74$, $p \leq .001$, $\mu_p^2 = .66$
Difficulty	2	$F(2,36) = 14.30$, $p \leq .001$, $\mu_p^2 = .44$	3	$F(2,24) = 12.68$, $p \leq .001$, $\mu_p^2 = .51$
	4	$F(2,24) = 10.37$, $p \leq .001$, $\mu_p^2 = .46$		
Confidence Required	3	$F(1,12) = 9.13$, $p \leq .011$, $\mu_p^2 = .43$		
Instruction	4	$F(1,12) = 15.38$, $p \leq .002$, $\mu_p^2 = .56$		

As noted in Table 1, nAFC had a consistent and reliable effect on RT, as did decisional difficulty. Predictably, more difficult decisions always took more time to make. Of the 16 possible RT patterns generated by plotting RT as a function of nAFC at each level of difficulty for each experimental condition throughout the four experiments, 14 reflected the predictions made by Hick (1952), while the remaining 2 suggest linear increases in RT along with increases in decisional complexity (these were the plots generated for the "medium" and "hard" difficulty trials generated for the "regular instructions" condition of Experiment 4).

Participant RTs were reliably longer when post-decisional confidence rendering was required in Experiment 3, and participants took reliably longer to make decisions when instructions were convoluted (i.e. 2nd MOST/2nd LEAST) in Experiment 4 (Table 1).

Though not part of the formal analyses, it is interesting to note that participants generally took longer to decide "LESS" compared to "MORE", and longer to make incorrect decisions compared to correct decisions.

Mean Confidence Rating Analyses

Decisional difficulty had a reliable effect on mean confidence level, with difficult decisions resulting in lower mean confidence (Table 2). Decisional complexity also had a consistent and reliable effect of mean confidence. The mean confidence response patterns were most interesting: In all but one case (Experiment 1) participants exhibited a 'reverse Hick's Law' in their confidence judgements. They grew less confident as decisional complexity increased, but the differences in mean confidence decreased as decisional complexity increased.

Table 2. Mean confidence rating main effect analyses for each of the relevant experiments

Factor	Experiment number and Result		Experiment number and Result	
nAFC	1	$F(2,36) = 8.10, p \leq .001, \mu_p^2 = .70$	2	$F(2,36) = 27.64, p \leq .001, \mu_p^2 = .61$
	3	$F(2,24) = 18.33, p \leq .001, \mu_p^2 = .60$	4	$F(2,24) = 49.07, p \leq .001, \mu_p^2 = .80$
Difficulty	2	$F(2,36) = 19.05, p \leq .001, \mu_p^2 = .51$	3	$F(2,24) = 17.98, p \leq .001, \mu_p^2 = .60$
	4	$F(2,24) = 41.50, p \leq .001, \mu_p^2 = .78$		
Instruction	4	$F(1,12) = 11.64, p \leq .005, \mu_p^2 = .49$		

Decisional Accuracy Analyses

Predictably, participants became less accurate as decisions became more complex and/or more difficult, and when the instructions became more convoluted (Table 3). Performance in all cases was better than chance.

Table 3. Decisional accuracy main effect analyses for each of the relevant experiments

Factor	Experiment number and Result		Experiment number and Result	
nAFC	1	$F(2,36) = 42.52, p \leq .001, \mu_p^2 = .70$	2	$F(2,36) = 237.38, p \leq .001, \mu_p^2 = .93$
	3	$F(2,24) = 371.49, p \leq .001, \mu_p^2 = .97$	4	$F(2,24) = 257.58, p \leq .001, \mu_p^2 = .96$
Difficulty	2	$F(2,36) = 110.10, p \leq .001, \mu_p^2 = .86$	3	$F(2,24) = 182.19, p \leq .001, \mu_p^2 = .94$
	4	$F(2,24) = 83.89, p \leq .001, \mu_p^2 = .88$		
Confidence Required	3	$\mathbf{F(1,12) = 3.88, p \leq .07, \mu_p^2 = .24}$*		
Instruction	4	$F(1,12) = 214.67, p \leq .001, \mu_p^2 = .95$		

*This result is not significant at $\alpha_{pc} = .05$.

Post-Decisional Mean Time to Render Confidence Analyses

CT only varied significantly within Experiment 4. CT was reliably less for "normal instruction" blocks relative to the "convoluted instruction" blocks [$F(1,12) = 4.81, p \leq .049, \mu_p^2 = .286$], and CT increased along with nAFC [$F(2,24) = 7.24, p \leq .003, \mu_p^2 = .38$].

Discussion

The RT and mean confidence response patterns generated by these four experiments are inconsistent with the idea that an ideal confidence level is used to regulate the amount of evidence accrued in a sensory-based decision-making task. Mean confidence did not remain constant with changes in decisional complexity, which suggests participants in these studies were not working to maintain an ideal level of confidence. Interestingly, decreases in confidence did seem to coincide with increases in decisional response times. This suggests either one is affecting the other, or that both are being affected by a similar process.

A review of the psychophysical literature yielded two possible explanations, and both were tested via a computer model.

The first possibility was suggested by Shannon (1951) who claimed the information gained following any single sampling event increases in a non-linear fashion as the number of possible event outcomes increases. Specifically, Shannon suggests $H(X) = -\Sigma p_i \log 2\ p_i$, where H(X) is the information obtained via sampling event X and pi is the probability of the event occurring. For the 'medium' difficulty decisions presented herein, where there were 1.08 'MOST' coloured squares for every 1 coloured square represented 2nd most, H(X) would have been .9988 for the 2AFC, 1.5822 for the 3AFC, and 1.9918 for the 4AFC. If one allows a constant rate of information processing, it should naturally take longer to process stimuli as decisional complexity increases, and RT patterns should become decidedly Hick-like.

Stevens' (1957) Power Law also suggests a reason for this non-linear evolution of RT. The well known theorem states $\Psi = k\Phi^a$, where Ψ is a psychological perception of the magnitude of stimulus intensity, k is an arbitrary scale unit, Φ is the actual physical magnitude of the perceived stimulus, and 'a' is the power exponent. Experiments with aversive stimuli, electric shocks for example, show that it takes increasingly smaller changes in the intensity of the physical magnitude of a stimulus to create changes in the psychological perception of stimulus intensity. In other words, as painful stimulus intensity increases linearly, the perception of pain increases exponentially. When exponent 'a' is assigned a positive value greater than 1, Stevens' Power Law describes this effect well.

If one allows that the complex decisions described above are more adverse than the simpler decisions (informal post-experiment interviews did indeed suggest participants did not enjoy the 4AFC decisions they were asked to make) then Stevens Power Law suggests an explanation for the Hick-like RT patterns, via a means of adjusting decisional criteria. Allow:

$$\text{Summed Discomfort} = \Sigma n^a \quad (1)$$

$$\text{Proportional Discomfort} = n^a / \text{Summed Discomfort} \quad (2)$$

$$\text{Criteria} = \text{Base Criteria} - \text{Proportional Discomfort} \quad (3)$$

where 'n' is the n from nAFC and 'a' is a Stevens-like exponent. For example, if Base Criteria were set to 5 (since Proportional Discomfort has a maximum value of 1, Base Criteria must necessarily be small) and exponent 'a' were set to 10, correct decision E(RT) values for the 'medium' difficulty decisions in these experiments would be: E(RT|2AFC) = 9.5976 time units, E(RT|3AFC) = 13.7408 time units, and E(RT|4AFC) = 17.8206 time units, yielding a slight Hick effect.

Computer based evidence accrual models were designed in MATLAB and used to test both of these suggestions. The best fit to the general trends in the observed data came from a combination of both. As can be seen from a comparison of Figure 1 panels 1 and 2, a reasonable approximation to the data was obtained when total RT within the model was allowed to increase at each evidence accrual event by the amount of information Shannon suggests would be obtained (mimicking a constant rate of evidence accrual through time),

decisional criterion levels were adjusted as described in (5) above, and mean confidence was allowed to equal the proportion of accrued evidence which supported the ultimate decision (as an aside: The model also required a minimal amount of confidence related evidence accrual, which allowed for post-decisional confidence processing when the model was asked to simulate decisions made under speed stress). The data provided in Figure 1 are based on a simulated 1000 decisions per nAFC level per level of decisional difficulty, with Base Criteria set to 5 (with each evidence accrual event involving a simulated sampling of a 5x5 segment of the coloured stimuli), exponent 'a' set to 10, and RT set to increase with each accrual by the information bit rate (as opposed to increasing at a constant rate). Though not depicted, it is also worth noting that modelled decisional accuracy very closely paralleled Experiment 2 decisional accuracy for each level of nAFC at each level of difficulty.

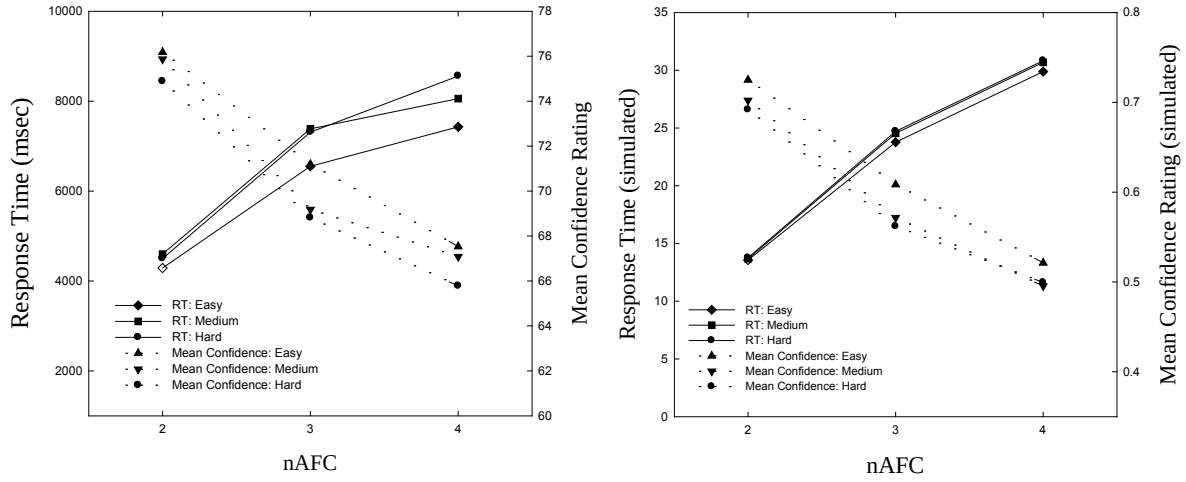

Fig. 1. RT and mean confidence for Experiment 2 (left panel) and the model (right panel).

Acknowledgements

This work was supported by a Natural Sciences and Engineering Research Council of Canada Graduate Scholarship to Carroll and Natural Sciences and Engineering Research Council of Canada Individual Discovery grant to Petrusic.

References

Hick, W. E. (1952). On the rate of gain of information. *Quarterly Journal of Experimental Psychology*, 4, 11-26.

Petrusic, W. M. (1992). Semantic congruity effects and theories of the comparison process. *Human Perception and Performance*, 18 (4), 962-986.

Shannon, C. E. (1951). Prediction and entropy of printed english. *The Bell System Technical Journal*, 30, 50-64.

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review*, 64, 153-181.

Van Zandt, T. (2000). ROC curves and confidence judgements in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26, 582-600.

Vickers, D. (1979). *Decision processes in visual perception*. New York: Academic Press, Inc.

Vickers, D. & Lee, M. D. (1998). Dynamic models of simple judgments: I. properties of a self-regulating accumulator module. *Nonlinear Dynamics, Psychology, and Life Sciences.* 2 (3), 169-194.