The Dynamics of Experimentally Induced Criterion Shifts

Scott Brown

University of California, Irvine

Mark Steyvers

University of California, Irvine

Address for correspondence:

Scott Brown

Department of Cognitive Sciences

3151 Social Sciences Plaza

University of California

Irvine, CA 92697-5100

Ph. (949) 824 2051

Email: scottb@uci.edu

Abstract

Investigations of decision-making have typically assumed stationarity, even though commonly observed “context effects” are dynamic by definition. Mirror effects are an important class of context effects that can be explained by changes in particpants’ decision criteria. When easy and difficult conditions are blocked alternately and a mirror effect is observed, participants must repeatedly change their decision criteria. We investigate the time course of these criterion changes, and observe the build up of mirror effects on a trial-by-trial basis. Our data are consistent with slow, systematic changes in decision criteria that lag behind stimulus changes. The length of this lag is considerable: analysis of a simple dynamic signal detection model suggests participants take an average of around 14 trials to adjust to new decision environments. This trial-level measurement of experimentally induced changes in criterion has implications for traditional block-wise analyses of data, and for models of decision-making.

A common assumption in models of decision-making is *stationarity*. With few exceptions (e.g., Kac, 1966; Rabbit, 1981; Strayer & Kramer, 1994a,b; Treisman & Williams, 1984; Vickers & Lee, 1988, 2000), models of decision-making assume that successive decisions are independent. The assumption of stationarity has proven useful in keeping models simple and tractable, and seems reasonable as most decision-making experiments have employed stationary decision-making environments. More recently, there has been a growing focus on non-stationary (dynamic) research. A central feature of most dynamic research in psychology is a focus on behavioral changes triggered by internal events, such as stimulus or response monitoring, or error-rate tracking (e.g., Heit, Brockdorff, & Lambert, 2003; Kelly, Heath & Longstaff, 2001; Petrov & Anderson, in press; Rotello & Heit, 2000; Triesman & Williams, 1984; Van Orden, Moreno & Holden, 2003). Often, these internally induced changes are fast, on the order of seconds (although see also Gilden, Thornton, Mallon, 1995). The key aspect of internally induced changes is that they can occur at any point during measurement – there is no way to predict their arrival times before the experiment begins.

Below, we consider decision environments that are themselves dynamic, experimentally inducing changes in behavior. For example, consider the case of a medical observer making decisions about the nature of tumors (benign vs. malignant) from x-ray photographs. Decision difficulty will change with time, as the patient population, or even the picture clarity changes. Observers must dynamically adjust their decision-making processes to reflect changes in the environment: if it becomes easier to identify benign tumors, observers should relax their criterion for identifying malignant tumors. In this article, we report an empirical and theoretical investigation of this classic

criterion setting problem. We introduce a simple decision model based on signal detection theory (SDT) to measure changes in criterion, and fit this model to data from four experiments in which we experimentally induce changes in the decision criterion.

Our research addresses the dynamics of criterion shifts induced by experimental manipulations. These manipulations lead to simple a priori predictions for the timing of the induced criterion changes. Experimental manipulations set up a dynamic decision-making environment in which the predicted criterion changes can be measured. Any kind of stimuli could be used for these environments; for generality we refer to the two classes of stimuli as “targets” and “distractors”. These stimuli could be benign versus malignant tumors, or words versus nonwords in a lexical decision task, or any of a host of other examples. We define two different decision environments by the properties of their distractors. In one environment, the distractors may be relatively dissimilar from the targets, making decisions relatively easy. In the other, the two types may be much more alike, resulting in relatively hard decisions. We then construct a dynamic decision environment by alternating sequences of easy and hard decisions, as shown in Figure 1.

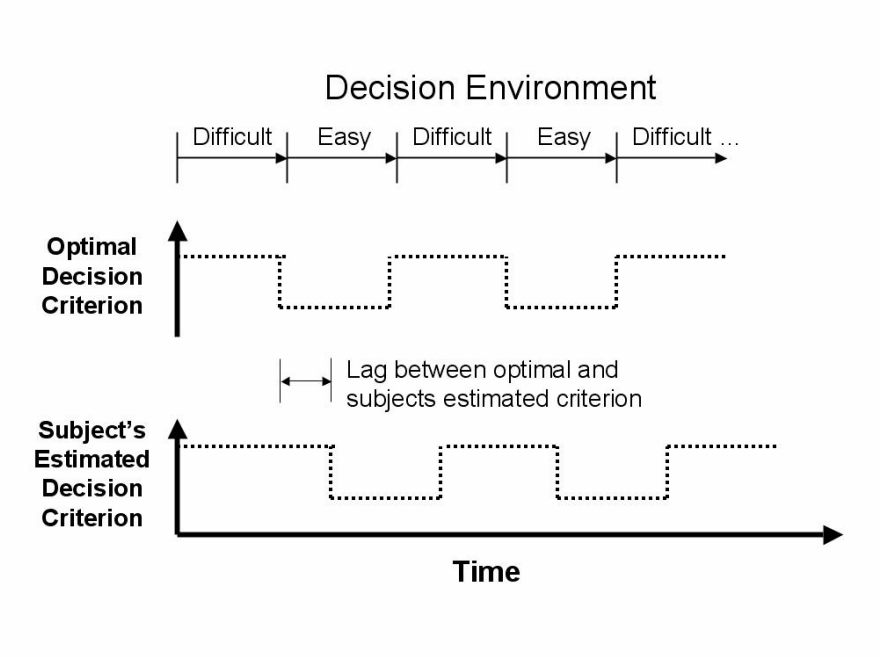

Figure 1: Basic paradigm. Decision environments change (top) leading to changes in the optimal decision criterion (middle). Participants' actual decision-making processes (bottom) lag behind.

The alternating decision contexts described in Figure 1 have been widely used in *blocked* cognitive psychology experiments, often leading to *context* or *blocking* effects. A context effect occurs when behavior associated with an experimental condition is different at different times – even though the condition itself is unchanged – because the context of the condition has changed. One particularly prominent context effect is the *mirror effect*, describing a particular relationship between performance levels in a pair of decision environments of different difficulty. A mirror effect is said to occur when performance in the easier condition is marked by better performance on *both* of the response alternatives. For example, in recognition memory, a mirror effect occurs when the condition with higher accuracy has both higher hit rate (HR) and lower false alarm rate (FAR) than the condition with lower accuracy (e.g., see Glanzer, Adams, Iverson & Kim, 1993). A mirror effect can be observed when the properties of one stimulus type,

say distractor items, are changed. Changes in FAR are to be expected from changes to the stimuli with which they're associated (distractors) but mirror effects include changes in HR that cannot be explained this way. Since the properties of the target stimuli are unchanged, observed changes in HR must be due to changes in participants' decision-making processes.

Our focus is on the dynamic properties of mirror effects – how they are established over time, and how they change when decision-making contexts are changed. Mirror effects are most conveniently explained by changes in the location of a decision criterion between the high and low accuracy conditions, using SDT (Green & Swets, 1966). SDT posits that participants decide between two classes of items ("targets" and "distractors") by generating an internal magnitude for each stimulus and comparing that magnitude with a decision criterion, as illustrated in Figure 2. Target and distractor stimuli give rise distributions of internal magnitudes that cross over: some distractor stimuli have greater perceived magnitudes than some target stimuli, and vice versa. The decision criterion illustrated in the left panel of Figure 2 is optimal in the sense that it minimizes the total number of errors (misses + false alarms). We have drawn an optimal criterion for simplicity of display, but in model analyses we allow the criterion to be non-optimal (i.e., we estimate bias).

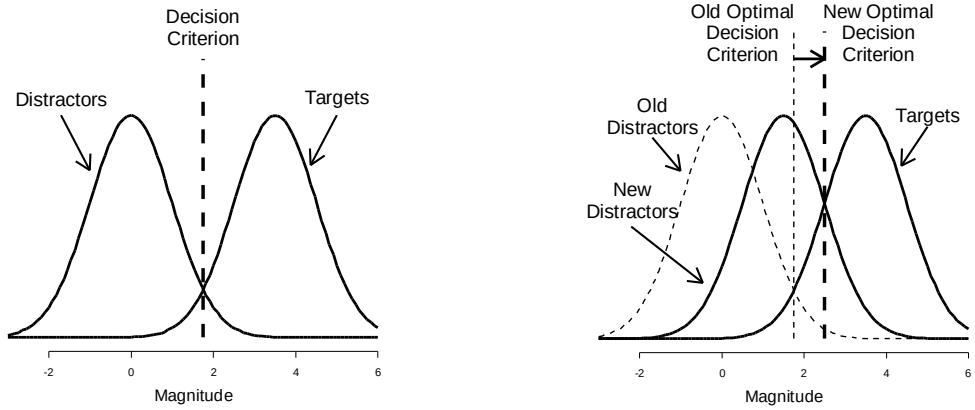

Figure 2: The left panel illustrates standard SDT: stimuli above the decision criterion are classified as targets; stimuli below are classified as distractors. The right panel shows how the optimal decision criterion changes when the properties of the distractors are altered.

In the SDT framework, a mirror effect can be caused by changing the properties of just the distractor items. Suppose the task becomes more difficult because the distractors are made more similar to the targets (shown in the right panel of Figure 2), then the old decision criterion is no longer optimal and must be raised. Intuitively, this represents the idea that participants recognize that moderate perceived magnitudes are now more likely than before to have come from the distractor distribution. Within the framework of SDT, a mirror effect due to changes in the properties of just the distractor items can *only* be explained by changes in the decision criterion: with an unchanged distribution for target items, observed changes in HR can *only* be due to changes in the decision criterion. While mirror effects can be explained by criterion shifts (e.g., Stretch & Wixted, 1998, but see also Mewhort & Johns, 2000), the time course of these shifts has been largely unstudied. Much research has assumed that criterion shifts occur in negligible time, but this is statistically impossible in many situations – some minimum number of samples from the new environment are required to identify a context change.

Below, we present experiments and theory investigating the time course of criterion shifts that establish mirror effects.

Simple Dynamic Measurement Model

We have developed a simple dynamic version of SDT to approximate the expected behavior of an observer in the dynamic decision-making paradigm outlined above (in Figure 1). In its simplest form, the dynamic SDT model is designed to apply to decision-making tasks where there are two different decision-making environments that alternate throughout the task. The model is based on two static SDT models, one for each decision-making environment, where one environment is more difficult than the other. The model assumes that there is an SDT model operating in the difficult environment, defined by a sensitivity parameter (d'_H, "H" for "hard") and a decision criterion (C_H) and another SDT model operating in the easy decision environment, defined by d'_E and C_E ("E" for "easy"). We assume equal variances for the target and distractor distributions, although later we discuss – and estimate – the unequal variance case.

The crucial addition that allows us to model dynamic behavior is that we assume that the criterion *lags* when decision environments change. For example, when the decision environment changes from easy to difficult, we assume that the sensitivity of decisions changes immediately, from d'_E to d'_H. Immediacy makes sense given that only the stimuli themselves define decision difficulty. By contrast, the decision criterion is under the control of the decision-maker, and thus will not change until they notice the change in decision environment, or some correlated variable (e.g., changed error rates). In our example, when changing from an easy to a hard decision environment, we assume that the decision criterion only changes from C_E to C_H after some lag, L. This assumption

of a step-wise change in criterion may seem overly simple, and is different from the incremental adjustments assumed by others (e.g., Strayer & Kramer, 1994b; Treisman & Williams, 1984). We examined other assumptions, such as a smooth exponential approach from the old to the new criterion, or a piecewise-linear approach, and found that they provided no significant improvement in fit. Given that the data could not discriminate between the various possibilities, we chose the stepwise criterion change for its computational simplicity and the interpretability of its parameter *L* (the number of decisions after an environment change before the participant changes their criterion).

Model Predictions

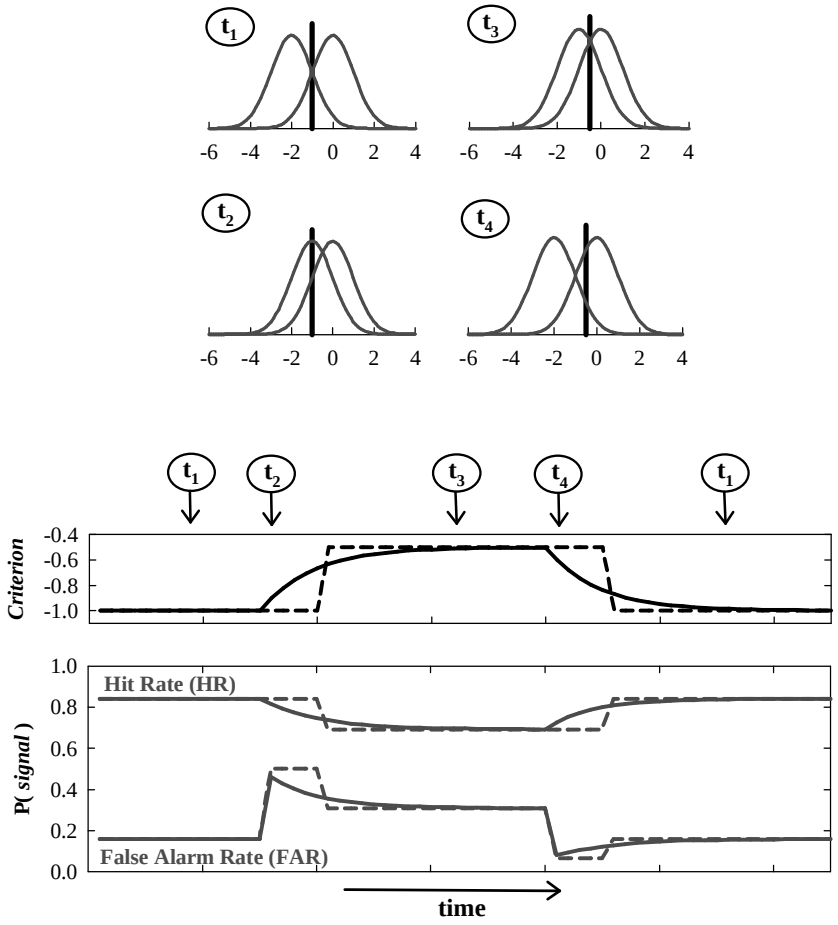

Figure 3: Predictions from the dynamic SDT model. Upper part shows static SDT submodels. Lower part shows predicted criterion changes (top) and predicted hit and false alarrm rate changes (bottom).

Figure 3 makes the predictions of this model clearer. Once again, we have used optimal decision criteria on Figure 3, for simplicity of illustration. When fitting the model to data, we allow for non-optimal criteria by estimating bias parameters. The SDT model illustrated in the top-left corner, t_1, illustrates behavior during easy decisions: d'_E is relatively large (the signal and noise distributions are relatively far apart) and the criterion C_E is approximately optimal. This submodel leads to the hit rate (HR) and false alarm rate (FAR) predictions at the left edge of the right-hand panel, with high HR and low FAR. Suppose the decision environment changes from easy to hard at time t_2, when the distractor stimuli become more similar to the targets. The SDT model then operating is shown in the lower left corner, with label t_2: note that sensitivity has decreased due to the harder stimuli (d'_E has changed to d'_H), but the criterion has not yet changed. This leads to the HR and FAR predictions shown under t_2 in the lower right-hand panel in dashed lines: no immediate change in HR, but a large increase in FAR. After some lag, L, the decision-maker updates their criterion to C_H, the criterion for hard decision environments (shown by the dashed line in the upper left plot). The SDT model then operating is shown as t_3 on the left hand side of Figure 3, and its predictions are shown by the dashed lines under the label t_3 in the lower right-hand plot: a decrease in both HR and FAR. Finally, the decision environment again changes, back to the easy condition, changing sensitivity but not immediately changing the decision criterion. This corresponds to SDT model t_4 and a predicted decrease in FAR, with no change in HR. Again, after some lag, the decision criterion is changed to C_E, bringing us back to the SDT model t_1.

The dashed lines in Figure 3 show predictions for an individual subject: our assumption of stepwise criterion changes results in stepwise changes in predicted HR and FAR. When analyzing data below, we show fits to large groups of participants, where each participant is fit individually, but the observed and expected HR and FAR are averaged over subjects for graphing. Those graphs show smooth changes in FAR and HR, as illustrated by solid lines in the right-hand panels of Figure 3. Smooth transitions are the result of averaging over many individual stepwise transitions, with variable step positions. Note that the changes in FAR and HR for easy-to-hard and hard-to-easy transitions are symmetric. This is a consequence of the optimal criteria used in illustrating Figure 3, our use of non-optimal criteria in actual model fits allows for asymmetry in fits to data.

Equal vs. Unequal Variance Assumptions

The model we have outlined thus far assumes that the target and distractor distributions have equal variance, however we are not restricted to this equal variance assumption. In recognition memory tasks, many researchers have observed that the variance of the target distribution is greater than that of the distractor distribution (e.g., Glanzer, Kim, Hilford, & Adams, 1999; Malmberg, 2002; Ratcliff, Sheu, & Gronlund, 1992; Sheu & Heathcote, 2002; Verde & Rotello, 2003). In recognition memory studies, the relative variances of the two distributions is usually estimated by constructing receiver operating characteristic (ROC) curves using data from confidence ratings. Below we detail an alternative method for estimating unequal variance models, essentially by estimating two-point ROC curves. Our methodology supports such estimates because our manipulation of decision difficulty causes changes in criterion

setting, as with confidence rating manipulation. More details are provided when discussing the model fits.

Change Mechanism

A limitation of the model so far is that it doesn't include a mechanism for how changes in criterion occur: what triggers a change? Mechanisms have been proposed based on response monitoring (Triesman and Williams, 1984), stimulus monitoring (Strayer & Kramer, 1994b; Vickers and Pietsch, 1998, 2000) and error rate monitoring (Rabbit, 1981). We use a simpler, descriptive model of criterion change that ignores smaller, spontaneous changes in criterion and focuses on the measurement of experimentally induced changes. The simplicity of this model also improves estimation reliability. Estimation reliability is particularly important, because dynamic measurement of experimentally induced criterion changes has not been previously attempted, yet has important implications for data analysis and model development.

Experiments 1-4

Mirror effects due to changes in decision difficulty have been observed across a wide range of decision tasks. For example, Stretch and Wixted (1998) identified mirror effects in episodic item recognition, when they manipulated memory strength by giving greater study opportunities for some items than others. Glanzer and Adams (1985, 1990) observed a similar effect when they manipulated recognition memory accuracy by changing the length of study lists. Mirror effects have also been observed in decision tasks other than recognition memory. In particular, robust context effects have been observed in lexical decision tasks, where participants classify strings of letters as words

(e.g. “CAT”) or nonwords (e.g., “CXT”) (see, e.g., Glanzer & Ehrenreich, 1979; Gordon, 1983; Grainger, 1996; Ratcliff, Gomez & McKoon, 2004).

Wagenmakers, Steyvers, Raaijmakers, Shiffrin, van Rijn, and Zeelenberg (in press) also identified a mirror effect in lexical decision. They observed improved performance for both words and nonwords (a mirror effect) when the similarity of the nonwords to the words was decreased. Mirror effects in lexical decision tasks usually include changes in both response time (RT) and accuracy. In our experiments, we use a variant of the signal to respond procedure, similar to that of Wagenmakers et al. and to Kello and Plaut’s (2000, 2003) “tempo naming task”. This procedure allows us to hold RT relatively constant, and so observe changes only in response accuracy, simplifying analysis and allowing comparison with the predictions of our dynamic SDT theory.

We employ a lexical decision task for Experiments 1-3 and a numerosity categorization task in Experiment 4. In each experiment, we alternated easy and difficult decision contexts. The easy and hard contexts always differed only in the properties of one of the stimulus classes – the properties of the other class remained unchanged throughout the experiment, allowing separation of effects due to criterion shifts from effects due to stimulus changes. In Experiments 1-3, the difficult decision context was defined by nonwords that were very similar to real words – they have a high “wordiness” – and the easy decision context was defined by nonwords with lower wordiness. In Experiment 1, the changes between hard and easy decision contexts occurred at random points within each block. Experiment 2 used a more typical “blocked” design, where the decision context only ever changed between blocks. Experiment 3 was the same as Experiment 2, except that participants were made aware of the experimental design

before beginning. In Experiment 4, subjects decided whether strings of arrows had more arrows facing leftwards or rightwards. The properties of one kind of display (left or right facing) were kept constant within-subject, while the properties of the other were manipulated to change decision difficulty. Changes in difficulty occurred only during block breaks.

General Methods

Procedure (Experiments 1-3)

The participants' task was always to respond with one mouse button if a letter string presented was a valid English word, and with another button if not (mouse button timing on our systems provides an accuracy of about +/-6ms, see Beringer, 1992). The buttons used for each response were counterbalanced across subjects. The same set of words and nonword strings were used in all three experiments. Seven letter words were drawn from the Kucera-Francis word pool, and nonwords were constructed by altering either just one letter of a valid word (making 'difficult' nonwords) or by altering three letters ('easy' nonwords), always checking to make sure that the resulting letter string was not a new valid word. The letters used to replace letters in valid words when creating nonwords were chosen from a multinomial distribution approximately matching the letter frequencies observed in written English: Table 1 gives examples of the stimuli.

Table 1. *Example words and easy and hard nonwords. Note that the "wordiness" of easy nonwords is lower than that of the hard nonwords.*

NONWORDS		**WORDS**
EASY	***HARD***	
CNOTSUN	SUBVIRT	PASSIVE
HASWEND	COMNLEX	DESCENT
FOMLERS	LIBFARY	CONICAL
BOEKLAW	PETWIFY	FURIOUS
EPPAASI	FROPLET	COMPOST
UNILIMA	PYRAMOD	FAILING
KTEDUAL	SUBJERT	ROYALTY
ROSTOMG	CINEGAR	INQUIRE
SEARAHE	KSOWING	PAINTER
REAYSED	CROQUIT	CURRANT

We used a variant of the signal-to-respond procedure in order to keep response latency as constant as possible, leaving accuracy as our only dependent variable, similar to Kello and Plaut's (2000, 2003) "tempo naming" task. In their procedure, a series of rhythmic tones are presented on each trial that help subjects anticipate the response signal. We generalized their method by keeping a constant rhythmic tone throughout each experimental block, continuing between trials. Every 400ms a 256Hz tone sounded for 50ms, and these beeps reliably indicated when stimuli would be presented, and also when responses were required. Each trial consisted of two beeps with a blank stimulus display area, followed by three "countdown" beeps during which the numbers "3", "2", "1" were displayed in the stimulus position. The stimulus character string was displayed on the beep immediately after the "1" and was removed from display on the following beep.

Participants were instructed to respond between 330ms and 700ms after the stimulus was displayed on the screen. If their responses were outside this window, they were given either "TOO FAST" or "TOO SLOW" feedback. To help subjects keep their responses within the acceptable window, a visible frame surrounding the stimuli changed

orientation during the “response window” time. This frame remained constantly visible throughout each block, changing only when a response was expected. Whenever there was no stimulus on display, the interior of the frame was blank.

At the end of each block participants were told their mean accuracy and response latency for that block, to help them maintain the desired performance level. As an extra aid all procedural timings slowed down by 50% in the first block (i.e., inter-beep time of 600ms), 33% in the second block (inter-beep time of 533ms) and 16.7% in the third block (inter-beep time of 467ms). All timings were rounded to the nearest integer multiple of the display monitor's vertical refresh period, which never resulted in a change of more than 7ms in any timing setting, and stimulus presentations were synchronized with the screen's vertical refresh.

Experiment 1 Details

Participants were 149 undergraduates from UC Irvine, who received course credit for participating. Data from subjects with an overall accuracy of less than 55%, or who had fewer than 70% of their responses within the acceptable latency window were discarded. This resulted in the loss of data from 14 participants (9.3%). Each participant in Experiment 1 completed 10 blocks of 100 trials each. Within each block, there was just one “switch point”, the position of which was distributed exponentially over trials greater than 20, with a mean switch point of trial 50 (the distribution truncated above trial 90). An exponential distribution of switch points was used for its constant hazard rate, making the probability of a switch occurring at any point, given that it had not previously occurred, constant. This makes the switch points least predictable, from a participant’s point of view.

At the switch point, the nonwords changed from either hard to easy or easy to hard. Changes from easy to hard nonwords always occurred on blocks after changes from hard to easy, and vice versa, so that stimulus properties were never changed between blocks. The switch point was constrained to be an even number, and there were always identical numbers of words and nonwords before the switch, and identical numbers of each after the switch point. Order of words and nonwords was selected by randomization without replacement, subject to the constraint that there were never more than five words or nonwords in succession. Participants were not informed about the changes between stimulus types, nor even that there were different classes of stimuli.

Experiment 2 Details

Experiment 2 had 108 undergraduate participants. Data from only two participants (1.85%) were rejected due to poor accuracy or inability to respond within the acceptable latency window. The improved performance over Experiment 1 most likely reflected the shorter blocks: there were 20 blocks of 40 trials each in Experiment 2. There were no "switch points" within blocks, so that blocked stimuli were always homogeneous. Each block had 20 words and 20 nonwords, ordered randomly, such that there were never more than five words or nonwords in succession. As before, the words used were always drawn from the same pool throughout the experiment, while the nonwords alternated from easy to difficult across blocks. The class of nonwords used for the first block (easy vs. difficult) was randomized across subjects. Participants were not informed about the classes of stimuli.

Experiment 3 Details

There were 169 undergraduate participants, with data from seven participants (4.1%) rejected due to poor performance. The design for Experiment 3 was identical to Experiment 2, except for the instructions given to participants. Participants were informed before the experiment began that there were two types of nonwords, those that were "easy to distinguish from real words" and those that were more difficult. They were shown examples of each class of nonwords. Before each block of trials began, a warning message was displayed informing participants what kind of block (easy or hard) was next. This warning was displayed in green for blocks with easy nonwords, and in red for blocks with difficult nonwords. During each block, the type of block (easy vs. difficult) was continuously displayed at the bottom of the display screen, in green or red.

Experiment 4 Details

We designed Experiment 4 to be conceptually similar to Experiment 2, while using a different choice task: numerosity instead of lexical decision. On each trial, participants in Experiment 4 were presented with a single row of 10 left and right pointing arrow symbols (two examples are shown in Figure 4). For each stimulus, the participants were to decide whether there were more arrows facing to the left or the right, and to push the corresponding mouse button. The left-to-right order of the arrows was randomly shuffled on every trial.

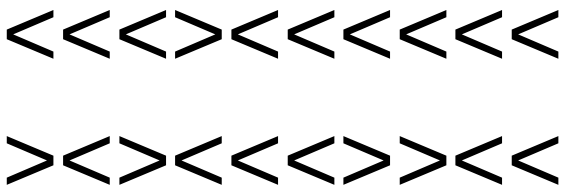

Figure 4: Example stimuli for Experiment 4. The appropriate response for each would be to push the *left* button, as more arrows face left than right. The upper stimulus is easier than the lower stimulus.

The same rhythmic beeping procedure was used as before and we manipulated task difficulty between blocks, as in Experiment 2. There were 20 blocks, each with 10 trials associated with "left" responses and 10 with "right" responses. Decision difficulty was manipulated by changing the distribution of proportions of left and right facing elements used: a display with 4/6 left/right elements (like the lower stimulus in Figure 4) is much more difficult than one with 1/9 left/right elements (like the upper stimulus in Figure 4). We always had right-favoring displays with either 4/6 or 3/7 proportions. We varied (between blocks) the left-facing displays from easy (either 9/1 or 8/2) to hard (either 7/3 or 6/4). We reversed the left/right assignment for half of the participants, although we collapse across this factor in all analyses below. There were 153 participants, and we discarded data from 24 participants who were unable to meet the accuracy and response deadline criteria.

Results

Experiment 1

We calculated the hit and false alarm rates for our participants, separately for blocks with easy and hard decision contexts. Figure 5 shows these data averaged across participants for Experiment 1.

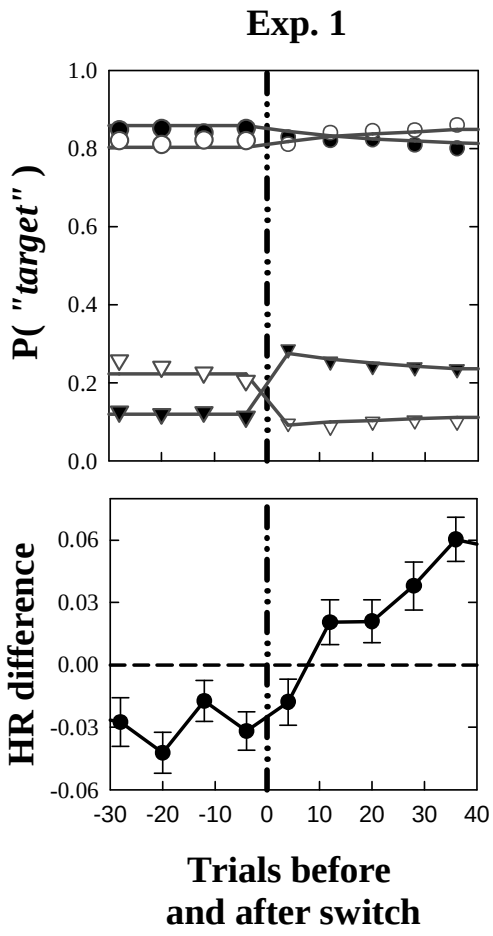

Figure 5: Data from Experiment 1, aligned to the stimulus switch point. In the top panel, circles represent HR and triangles FAR; filled symbols correspond to blocks where decision contexts changed from easy to hard, and open symbols represent blocks where decision changed from hard to easy. Bottom panel shows the difference between HR in easy and hard conditions changing within a block.

The data from Experiment 1 are shown in the upper panel of Figure 5, aligned at the switch point from easy to hard (filled symbols) or hard to easy (open symbols) decision contexts. The smooth lines are predicted probabilities from the dynamic SDT model, discussed later. The data were also averaged over blocks of eight consecutive trials for the purposes of graphing (although never for model estimation). The y-axis shows the probability of "word" responses. The FAR show that our manipulation of decision difficulty had the desired effect. When the decision context was easy (filled

symbols) the probability of incorrectly identifying a nonword as a word was low (FAR, triangles in Figure 5). When the nonwords were made more similar to words (after the switch point) the FAR jumped dramatically. A corresponding sudden decrease in FAR was observed when the nonwords were changed from difficult to easy (open symbols).

More interestingly, the change in nonword properties resulted in changes in the responses to *word* stimuli, shown by the hit rates (circles). After the nonwords changed from easier to more difficult (filled symbols) the HR steadily declined, and when the nonwords became easier, the HR steadily increased. These changes are consistent with our hypothesis of a lagged change in decision criterion. These HR changes describe a trial-by-trial emergence of the mirror effect. Before the switch point, there was a clear mirror effect: responses in the easy condition had both higher HR and lower FAR. Immediately after the stimulus switch point, the FAR reversed but there was no immediate change in the HR, and thus no mirror effect. With time, the HR reversed their ordering and thus the mirror effect re-emerged. This change took an average of about 12 trials after the stimulus switch, suggesting that there is a significant lag in participants' decision criterion changes. This is shown in greater detail in the lower panel of Figure 5, which plots the difference in hit rates between easy and hard conditions. Just before the stimulus switch point there was a reliable mirror effect: HRs in the easy condition were significantly higher than in the hard condition (one sample one-tailed $t(134)=3.4$, $p<.001$). In the eight-trial window following the switch point, the HRs were not significantly different ($t(134)=1.6$, $p>.05$). During trials 9-16, and thereafter, the HR for the easy condition was once again higher than for the hard condition (9-16 trials after

switch: $t(134)=1.9$; 17-24 trials after: $t(134)=2.1$; 25-32 trials after: $t(134)=3.3$; all $p<.05$).

The data from Experiment 1 show another interesting effect consistent with our lagged criterion change explanation. After the stimulus switch point, the FARs change suddenly and drastically, but then show a slow change back towards more intermediate levels. This pattern is also predicted by a lagged change in decision criteria: adjusting the decision criterion causes correlated changes in both HR and FAR.

Experiment 2 and 3 Results

Data from Experiments 2-4 are shown in Figure 6 in the same format as Figure 5.

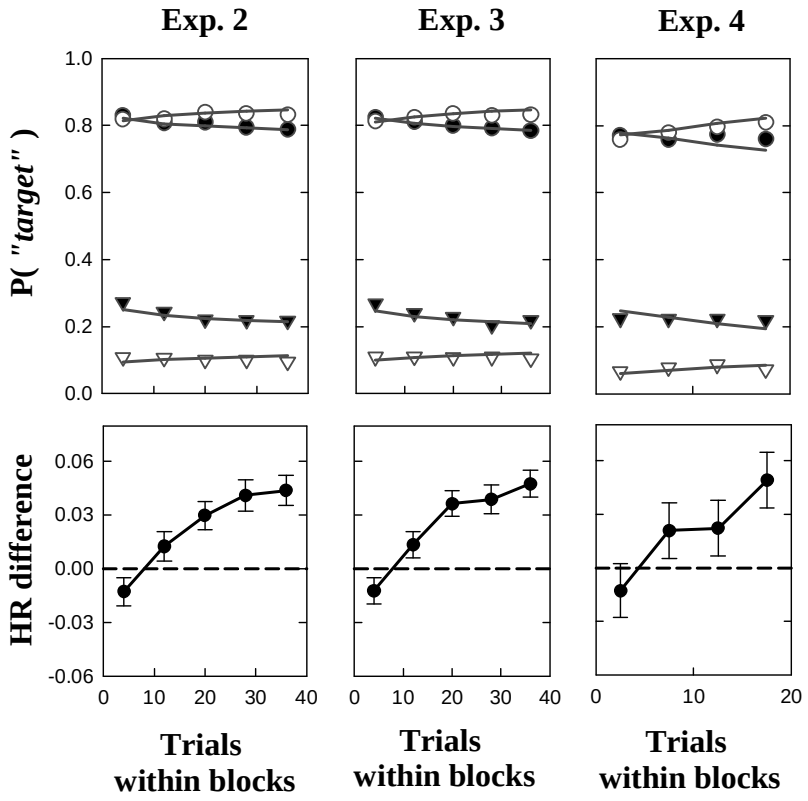

Figure 6: Data from Experiments 2-4, using similar format to Figure 5. In the top row: circles represent HR, triangles FAR; filled symbols correspond to hard decision contexts and open symbols correspond to easy decision environments. The bottom row of plots show the difference between hit rates in easy and hard conditions. The x-axes show trials within each block.

The data from Experiments 2 and 3 show similar patterns to those from Experiment 1. Recall that in Experiments 2-4 stimulus properties were only changed *between* blocks, so the effective stimulus switch point was trial zero, and data are aligned to that point for graphing. The FAR in blocks with easy decision contexts (open triangles) was much lower than in blocks with hard decision contexts (filled triangles). At the end of each block, presumably after changes in decision criterion had occurred, there were reliable mirror effects: HRs in the easy condition were significantly greater than in the hard condition (e.g., for trials 33-40 in Experiment 2, $t(105)=5.2$, $p<.001$, and in Experiment 3 $t(161)=6.4$, $p<.001$). There also appeared to be strong carry-over effects of task history: mirror effects were not present immediately after the stimulus switch points (block breaks). Hit rates in the easy condition were *smaller* than in the hard condition during the first eight trials of each block (in Experiment 2 the mean difference was 1.3%, $t(105)=1.6$, $p=.055$, in Experiment 3 mean difference was 1.2%, $t(161)=1.7$, $p<.05$). In Experiment 3 a statistically significant mirror effect re-emerged over trials 9-16 ($t(161)=1.8$, $p<.05$) and strengthened thereafter (trials 17-24, $t(161)=5.2$, $p<.001$). The changes were slower in Experiment 2: trials 9-16 showed a marginally significant mirror effect ($t(106)=1.5$, $p=.06$), which became significant over trials 17-24 ($t(106)=3.8$, $p<.001$).

The FAR from Experiment 3 demonstrate the same slow changes as observed in Experiment 1, consistent with our lagged criterion change explanation of these data. After the stimulus properties changed (between blocks) there was a sudden large change in FAR. The FAR then slowly changed over the course of the block towards more moderate values. These slow changes were always in the same direction as changes in

HR, consistent with a lagged criterion shift explanation. The changes in FAR from Experiment 2 were also as expected for the difficult blocks. However, for the easy blocks there was a very slight trend for the FAR to decrease across the block: 10.8% in trials 1-8, 10.7% in trials 9-16, 10.1% in trials 17-24, 10.0% in trials 25-32 and 9.5% in trials 33-40. These changes are not statistically significant (simple effects ANOVA: $F(4,101)<1$) but they are in the opposite direction to our predictions. They could presumably be modeled by within-block drifts in criterion placement or sensitivity. We do not pursue such models here, because the drift effects seem too small to justify the increased model complexity.

Other than the small drift effects, the results of Experiments 2 and 3 were very similar. The methodological difference between these experiments was in the information given to participants: in Experiment 2, participants were told as little as possible about the experimental design; in Experiment 3 they were fully informed and encouraged to switch between hard and easy decision environments as quickly as possible. The similarity of the data from these experiments suggests that participants are unable to adjust to new decision environments more quickly by recruiting conscious, intentional processes. Strayer and Kramer (1994a,b) found similar effects in their experiments: participants were unable to speed up adjustments of their (speed-accuracy tradeoff) criteria in response to instructions.

Experiment 4 Results

Experiment 4 used the same conceptual design as Experiments 2 and 3, but a different decision task: numerosity rather than lexical decision. The results of Experiment 4 are shown in the right-hand panels of Figure 6, averaging over blocks of

only five trials, rather than the previous eight, due to the smaller block lengths in this experiment. The data from Experiment 4 were very similar to those from Experiments 2 and 3. When the decision task was easy, the FAR (the chance of classifying a left-pointing stimulus as right-pointing – triangles in Figure 6) was much lower than when the task was difficult. As before, there was a reliable mirror effect in the latter part of each block: in trials 15-20 the mean difference was 4.9%, $t(128)=3.2$, $p<.001$. No mirror effect was observed in the first five trials of each block, the HR for the easy condition was *lower* than for the hard condition (by 1.3% on average). The HR ordering reversed during trials 6-10: easy HR was 2.1% higher than hard, but this was not yet reliable ($t(128)=1.3$, $p=.09$). A reliable mirror effect emerged during trials 11-15 ($t(128)=1.7$, $p<.05$).

In summary, the data from Experiments 1-4 all follow a simple pattern, with minor variations due to methodological changes. This pattern begins with a mirror effect: easy decision environments had both lower FAR and higher HR. When the decision context was made more difficult by changing *only* the properties of the distractor stimuli, there was a sudden change in FAR but no immediate change in HR. That is, the mirror effect was temporarily suspended when the distractor properties were altered. With time (an average of around 8-15 trials), the HRs changed to re-instate a mirror effect. Similar slow changes in FAR were observed, that were always in the same direction as changes in HR. All observed changes are qualitatively consistent with our dynamic SDT model that includes a lagged criterion shift.

Estimating the Dynamic SDT Model

To more accurately describe the data in terms of lagged criterion shifts, we estimated parameters for our dynamic SDT model, separately for each individual participant in each experiment. The model has five parameters: two sensitivities and two bias parameters to specify the *easy* and *hard* decision context SDT models (d'_E, d'_H, C_E and C_H) and a single *lag* parameter (L) that measures how many trials after stimulus properties are changed (d') before the decision criterion (C) is changed. We assumed that the decision criterion changed in a step-wise fashion between its easy and hard values, although this represents a computational convenience rather than a theoretical statement.

The HR and FAR probabilities can easily be calculated for each situation (easy and hard decision contexts, and lagged or not criterion values), conditional on parameter estimates. With these probabilities and the observed data it is simple to calculate maximum likelihood estimators of the model parameters by search[1]. The predicted hit and false alarm rates for this model are shown by the solid lines in the top panels of Figures 5 (for Experiment 1) and 6 (Experiments 2-4), aggregated in the same was as the data (within eight- or five-trial windows, and across participants). Histograms of the estimated lag parameters for all participants are shown in Figure 7.

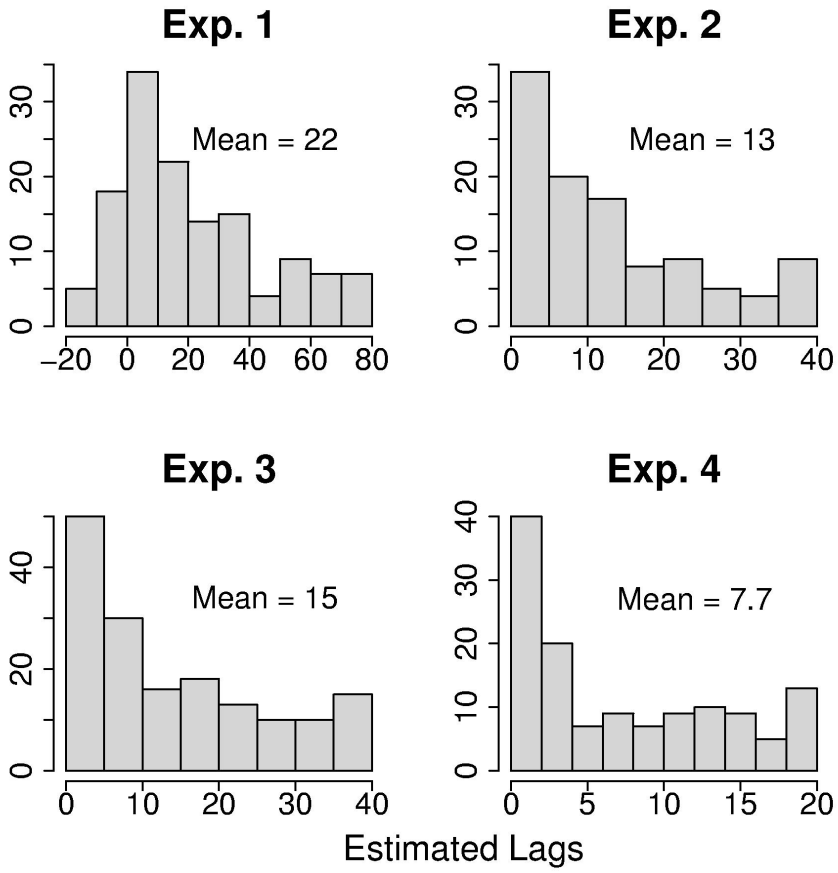

Figure 7: Histograms of estimated lags from fits of the dynamic SDT model.

The estimated lags show that most participants were quite good at appropriately changing their decision soon after the stimulus properties changed. In Experiment 1, 22% of participants had estimated lag parameters of between zero and five trials. Some participants behaved very differently: 8% had negative estimated lags, and 20% had estimated lags greater than 40 trials. Averaged over subjects, 22 trials (22% of block length) were needed to make the switch between contexts. Similar behavior, but without negative lags, was observed in Experiments 2-4 (recall that negative lags in Experiments 2-4 are isomorphic to long lags, as stimulus changes occurred in block breaks). Note the similarity of estimated lag values from Experiments 2 and 3 – again suggesting that

knowledge of the experimental design, and specific instructions to participants, does not decrease the amount of time taken to adjust to new decision environments (consistent with Strayer & Kramer, 1994a,b). In Experiments 2 and 3, averaged over subjects, 13 and 15 trials respectively were required for subjects to switch between contexts, representing 33% and 38% of block lengths. In Experiment 4, 7.7 trials were required on average, representing 39% of block lengths.

Unequal Variance Models

In addition to the above – equal variance – analyses, we have also investigated unequal variance models for our data. Most unequal variance SDT analyses are based on ROC curves derived from confidence ratings, typically in recognition memory paradigms. Although at least one attempt has been made to transport the ratings methodology to lexical decision (Jacobs, Graf & Kinder, 2003,) it relied on threshold presentation of stimuli. Therefore, the confidence ratings obtained may well have reflected confidence in perceptual processing rather in lexical decision processes.

Our methodology allows us to estimate unequal variance models without requiring confidence ratings. The key aspect of ratings tasks that allows estimation of variance parameters is that different confidence ratings are assumed to be generated by different decision criterion placements. Our manipulation of decision difficulty and the estimation of lagged criterion changes provides similar data. For example, in the easy decision condition, we are able to estimate hit and false alarm rates from decisions made after the criterion change (hence with a criterion appropriate to the easy context) and separately from decisions made before the criterion change (with a decision criterion appropriate to the difficult condition). Using these estimates, we were able to estimate

unequal variance models using standard maximum-likelihood techniques, simultaneous with estimation of the lag parameter. Data from Experiment 4 were not used for unequal variance estimation. In Experiment 4, the assignment of stimulus class to response type (target vs. distractor) was randomized across participants, as was the assignment of response type to response button. This symmetry makes the differentiation of "target" and "distractor" distributions purely formal, and makes unequal variance models implausible.

The first unequal variance SDT model we examined for Experiments 1-3 had one more parameter than the equal variance model (the ratio of variance in target distribution to that of the distractor distributions). This parameter did not significantly increase goodness-of-fit, as measured by χ^2 likelihood-ratio tests, for many participants (15%). The next unequal variance model we examined relaxed the assumption that the two distractor distributions (for easy and hard conditions) have the same variance. This assumption seemed unlikely, given that they have different mean values. We thus fit a model to the data in which the variance of each of the three (target and two distractor) distributions was related to their mean by a single parameter: $\sigma=\exp(-A\mu)$. This model includes an equal variance submodel ($A=0$), submodels in which target distributions have higher variance that distractor distributions ($A<0$), or lower variance ($A>0$). This model also did not provide a better fit to the data of very many participants (17%).

It is possible that the estimation of unequal variance models was numerically problematic. Estimation of the unequal variance parameters relies on separating data from before and after the estimated criterion switch point. When short lags were estimated, there was little data for this estimation, so numerical difficulties could have

resulted in non-optimal fits. As a check against this, we fixed the variance parameters at across all subjects and estimated the other parameters, just like estimating the equal-variance model. We performed this analysis for many different unequal variance parameters. These models have no more parameters than the equal variance model, and so standard model selection techniques (AIC, BIC, etc.) suggest simple selection based on likelihood value only. The very best performance we observed for these "fixed" unequal variance models was for a model in which the target distribution had unit variance, the hard distractors had variance $1/\sqrt{2}$ and the easy distractors had variance 1/2. This model had higher likelihood for 67% of the participants. While this is significantly different from the 50% expected by chance, it was not overwhelming support for unequal variance models. Further, the mean increase in likelihood was very small (less than 0.2%).

Even though the unequal variance models did not fit the data significantly better than the equal variance model, it is possible that they resulted in different parameter estimates, particularly criterion lag estimates. We tested this by comparing lag parameter estimates from the equal variance and unequal variance model fits using two-tailed repeated-measures t-tests, for experiments 1-3. We used parameter estimates from the simplest unequal variance model (the first one detailed above), reasoning that those estimates would be most reliable. In each of the three experiments, there was no significant difference between lag estimates under equal and unequal variance models: Experiment 1 mean difference of 0.03 trials ($t(134)<1$); Experiment 2 mean difference of 2.4 trials ($t(105)=1.7$, $p>.05$); Experiment 3 mean difference of 0.2 trials ($t(161)<1$).

Summary of Results

The experiments and data analyses presented above demonstrate the dynamic build-up of mirror effects over time. Mirror effects were established by changing the difficulty of decisions, and hit and false alarm rate changes were observed following changes in decision difficulty. We developed a simple dynamic version of SDT in which the decision criterion changes some time (*lag*) after stimulus properties change, and we used this model to fit data at an individual subject and individual trial level. Model-based analyses estimated the time required to adjust to new decision environments as around 14 trials, on average. This implies that a significant amount of data from "hard" (respectively, "easy") decision contexts actually reflects participants' "easy" (resp. "hard") performance mode, possibly contaminating typical data analyses, in which such dynamic changes are not taken into account.

Further analysis of the mirror effect magnitude demonstrates that this contamination could result in the reduction of mirror effect size by about 10% if data are subjected to the usual block-wise analyses. Using data from Experiments 2 and 3 (most similar to standard designs) we estimated the size of the mirror effect by calculating the mean difference in HR between the easy and hard conditions across participants, and dividing by the standard deviation of those differences to create a normalized effect size. The standard block-wise analysis (without excluding any data) showed a mirror effect size of 0.63 standard deviations in Experiment 2, and 0.57 in Experiment 3. We then excluded data from the first *N* trials of each block, choosing *N* to maximize the observed effect size. This involved a trade-off between increasing hit rate differences and increasing variability due to decreasing sample sizes. For Experiment 2, the maximum effect size was 0.70 standard deviations (an increase of 11%) occurring when we

removed data from the first four trials of each block. For Experiment 3, the maximum effect size was 0.63 standard deviations (an increase 10%) occurring when we removed the first five trials of each block.

General Discussion

The blocking paradigm described in Figure 1 experimentally manipulates the position of the optimal decision criterion, and thereby *induces* changes in participants' decision criteria. This manipulation entails strong constraints on hypotheses about exactly when we should observe criterion shifts, and what those shifts should look like, departing from previous work in some important ways (but see also Strayer & Kramer, 1994b). A *stationary* experimental design is one in which the properties of the task do not change during the experiment, hence participants do not need to change their behavior during the experiment in order to remain optimal. Experiments with between-subjects designs are typical of this category – the participants' task does not change during the experiment, so there is no compelling reason to consider sequential effects. Because static experiments are limited in their design researchers often use *dynamic* experiments, meaning that experimental conditions change with time, forcing participants to adjust their decision-making processes in order to remain optimal. Research with dynamic experimental designs but static analyses is common in psychology: blocked designs are used, making the task dynamic, but static analyses are applied because researchers (often implicitly) assume that sequential dependencies between blocks are either unimportant or unmeasurable. Some research does employ dynamic analyses, but mostly using static experiments in which dynamic behavior arises spontaneously, without being required by design. Most research into the presence of short-term autocorrelations, or of chaos and

longer-term nonlinear dynamics in behavioral data is of this kind (e.g., Gilden et al., 1995; Kelly et al., 2001; Van Orden et al., 2003) including previous examinations of the criterion setting problem (e.g., Kac, 1966; Rabbit, 1981; Treisman & Williams, 1984).

Our work so far only describes the time course of criterion shifts, without addressing the question: what *causes* the shifts to occur? There are a multitude of plausible theories to explain how participants adjust their decision criteria. The dynamic version of SDT we use is similar to Treisman and Williams' (1984) theory for criterion setting. In their theory, participants are assumed to change their decision criterion after each trial based on the stimuli and responses from the past few trials. Our dynamic SDT allows for a *lag* parameter that measures how far behind stimulus changes participants change their decision criteria, and is thus a simplified version of Treisman and Williams' model. The value of Treisman and Williams' more complex model is that it provides an account of trial-by-trial sequential effects in criterion setting, and it provides a mechanism by which criterion change comes about (response monitoring). For example, the obvious extension of Treisman and Williams' (1984) criterion setting theory would allow for slow adjustments caused by response monitoring (see also Rabbit, 1981). A discrete switching model may posit that participants estimate properties of the decision environment, and discretely switch between one set of assumed properties and another only when evidence against the status quo reaches some critical level. All of these classes of models are interesting explanations of the processes underlying decision criterion setting. However, at a first attempt, our dynamic SDT model is sufficient to *describe* the data, and provide useful measurements. Our simpler model affords important advantages in descriptive power and parameter estimation, allowing accurate estimation

of parameters for individual participants. It is also consistent with each of the model types just mentioned: at a sufficiently general level, each can be reduced to a two-state model in which changes in task properties precede changes in behavior.

Our observation of slow and systematic transient effects in criterion setting presents a challenge to previous modeling exercises that are static (i.e., do not include effects of stimlus history). For example, Ratcliff et al. (2004) identify context effects in lexical decision very similar to those in our Experiments 1-3. Ratcliff et al. model these effects using a diffusion model account of response time and accuracy, where the parameters that encode stimulus properties ("drift rates") are assumed to be different for the different stimulus classes (e.g., easy nonwords, difficult nonwords and words). This assumption is equivalent to our assumption that the signal and noise distributions in our SDT change with changing stimulus properties. However, Ratcliff et al. also implicitly model *context effects* with their drift rates. For example, they observe a context effect in which the differences in responses to different word classes are smaller when the nonwords were random letter strings than pseudowords; this effect is captured in Ratcliff et al.'s model by different drift rate parameters for words in the context of the two different kinds of nonwords. While this method of modeling the data is appropriate for Ratcliff et al.'s purposes, it neglects the fact that context effects must build up slowly.

Strategy Or Criterion Switch: Equivalent Models?

Some readers may wonder why we have chosen to model the difference in behavior from easy and hard decision contexts as a criterion shift, rather than a strategy shift. In fact, at the general, descriptive level of our SDT model, the difference is immaterial (see Ratcliff et al., 2004, for a similar argument). For example, suppose that

the decisions in question were made using one or other of two decision "modules", one module best suited to use in easy decision contexts one best suited to difficult contexts, and that the switch between usage of these modules lags behind the switch in stimulus properties. For the lexical decision task we used in Experiments 1-3, the "hard" module may involve determining whether the stimulus exists in the lexicon (a slower but reliable strategy) and the "easy" module may involve assessment of a word's "familiarity" (a faster but less reliable strategy).

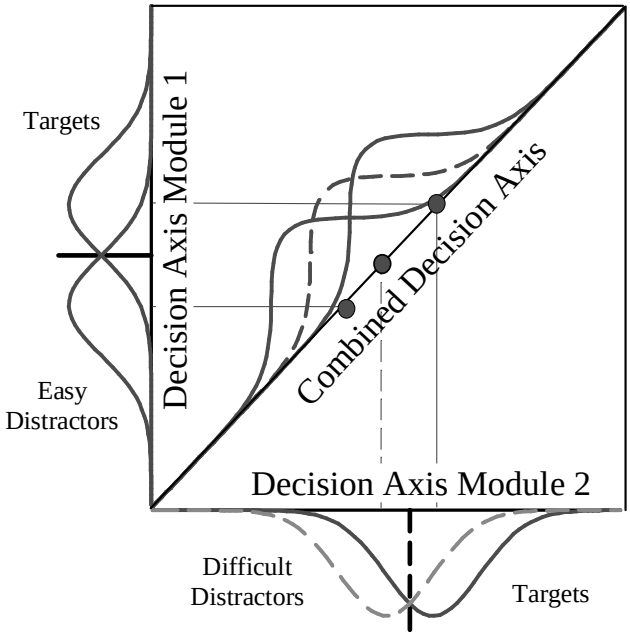

Figure 8: Isomorphism (at a descriptive level) of simple strategy shift and criterion shift models.

When there is a response deadline, it is reasonable to assume that the two decision modules produce response probabilities in a similar fashion to SDT. With these assumptions, the strategy switching model is isomorphic to our criterion shifting SDT model, as illustrated in Figure 8. The diagonal line shows the location of the means of the target distribution (upper right) and easy and hard distractor distributions (lower left

and middle, respectively). As with the dynamic SDT model, these distributions are assumed to change exactly when the stimulus properties change. We assume that each decision module produces distributions of some decision variable: the projection of the stimulus distributions onto the y-axis shows these distributions when using Decision Module 1, the module appropriate for easy decisions; the projection onto the x-axis shows the distributions under Decision Module 2. These distributions are the same as the non-lagged "easy" and "hard" SDT model subcases. Finally, if a strategy shifting lag is introduced so that, just after the stimulus properties change, the "wrong" decision module is employed for a short time, the outputs from this model are the same as those from our lagged SDT model.

Conclusions

When the properties of decision-making tasks change during experiments, participants' behavior must lag behind these changes. Our experiments show that this lag can be considerable in the case of alternating easy and difficult decision environments, so that behavior in each environment is influenced by the previous environment for many trials. We show that these effects are both qualitatively and quantitatively consistent with a simple dynamic version of SDT in which changes in decision criterion lag behind stimulus changes. These lags could have consequences for data analysis techniques, and for model development in decision-making paradigms. Realistic decision-making environments are likely to be much more variable than experimental ones, and so dynamic effects in real decision-making tasks may be very important, and are certainly poorly understood.

References

Beringer, J. (1992). Timing accuracy of mouse response registration on the IBM microcomputer family. *Behavior Research, Methods, Instruments and Computers, 24*(3), 486-490.

Gilden, D. L., Thornton, T., & Mallon, M. W. (1995). 1/f noise in human cognition. *Science, 267,* 1837-1839.

Glanzer, M. & Adams, J.K. (1985). The mirror effect in recognition memory. *Memory & Cognition, 13*(1), 8-20

Glanzer, M. & Adams, J.K. (1990). The mirror effect in recognition memory: Data and theory. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 16*(1), 5-16

Glanzer, M., Adams, J.K., Iverson, G.J. & Kim, K. (1993). The regularities of recognition memory. *Psychological Review, 100,* 546-567.

Glanzer, M. & Ehrenreich, S.L. (1979). Structure and search of the internal lexicon. *Journal of Verbal Learning and Verbal Behavior, 18,* 381-398

Glanzer, M., Kim, K., Hilford, A., & Adams, J.K. (1999). Slope of the receiver-operating characteristic in recognition memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 25*(2) 500-513

Gordon, B. (1983). Lexical structure and lexical decision: Mechanisms of frequency sensitivity. *Journal of Verbal Learning and Verbal Behavior, 22,* 24-44

Grainger, J. & Jacobs, L. (1996). Orthographic processing in visual word recognition: A multiple read-out model. *Psychological Review, 103*(3), 518-565.

Green, D. M. & Swets, J. A. (1966). *Signal Detection Theory and Psychophysics,* New York: John Wiley and Sons, Inc.

Heit, E., Brockdorff, N., & Lamberts, K. (2003). Adaptive changes of response criterion in recognition memory. *Psychonomic Bulletin & Review, 10*(3), 718-723

Jacobs, A.M., Graf, R., & Kinder, A. (2003). Receiver-operating characteristics in the lexical decision task: Evidence for a simple signal detection process simulated by the multiple-readout model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*(3), 481–488.

Kac, M. (1966). Some mathematical models in science. *Science, 166*(7), 695-699

Kello, C.T. & Plaut, D.C. (2003). Strategic control over rate of processing in word reading: A computational investigation. *Journal of Memory and Language, 48,* 207-232

Kello, C.T. & Plaut, D.C. (2000). Strategic control in word reading: Evidence from speeded responding in the tempo-naming task. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *26*(3), 719-750.

Kelly, A., Heath, R.A. & Longstaff, M. (2001). Response-time dynamics: Evidence for linear and low-dimensional nonlinear structure in human choice sequence. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology. Vol 55a*(3), 805-840

Malmberg, K. J. (2002). On the form of ROCs constructed from confidence ratings. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 28(2), 380-387. Meissner, C.A. & Brigham, J.C. (2001). Thirty years of investigating the own-

race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, & Law, 7*(1), 3-35

Mewhort, D.J.K. & Johns, E.E. (2000). The extra-list feature effect: A test of item matching in short term recognition memory. *Journal of Experimental Psychology: General, 129,* 262-284

Petrov, A., & Anderson, J..R. (in press). The Dynamics of Scaling: A Memory-Based Anchor Model of Category Rating and Absolute Identification. *Psychological Review.*

Rabbit, P. (1981). Sequential reactions. In D.H. Holding (Ed.) *Human skills.* pp. 153-175. London: Wiley.

Ratcliff, R., Gomez, P. & McKoon, G. (2004). A diffusion model account of the lexical decision task. *Psychological Review, 111*(1) 159-182

Ratcliff, R., Sheu, C.-F., & Gronlund, S.D. (1992). Testing global memory models using ROC curves. *Psychological Review, 99,* 518-535.

Rotello, C.M., & Heit, E. (2000). Associative recognition: A case of recall-to-reject processing. *Memory & Cognition, 28,* 907-922.

Sheu, C-F., & Heathcote, A. (2002). A nonlinear regression approach to estimating signal detection models for rating data. *Behavior Research Methods, Instruments & Computers, 33,* 108-114.

Strayer, D.L. & Kramer, A.F. (1994a). Strategies and Automaticity: I. Basic Findings and Conceptual Framework. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(2), 318-341

Strayer, D. L. & Kramer, A. F. (1994b). Strategies and automaticity II: Dynamic aspects of strategy adjustment. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 342-365

Stretch, V. & Wixted, J.T. (1998). On the difference between strength-based and frequency-based mirror effects in recognition memory. *Journal of Experimental Psychology: Learning, Memory and Cognition, 24*(6) 1379-1396

Treisman, M. & Williams, T.C. (1984). A theory of criterion setting with an application to sequential dependencies. *Psychological Review, 91*(1), 68-111

Van Orden, G.C., Holden, J.G., Turvey, M.T. (2003). Self organization of cognitive performance. *Journal of Experimental Psychology: General, 132*(3) 331-350

Verde, M.F., & Rotello, C.M. (2003). Does familiarity change in the revelation effect? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*(5) 739-746.

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

Vickers, D. & Lee, M.D. (2000). Dynamic models of simple judgments: II. Properties of a self-organizing PAGAN (Parallel, Adaptive, Generalized Accumulator Network) model for multi-choice tasks. *Nonlinear Dynamics, Psychology, and Life Sciences, 4*(1), 1-31

Wagenmakers, E.J.M., Steyvers, M., Raaijmakers, J.G.W., Shiffrin, R.M., van Rijn, H., & Zeelenberg, R. (in press). A Model for Evidence Accumulation in the Lexical Decision Task. *Cognitive Psychology*.

Acknowledgements

We'd like to thank EJ Wagenmakers, Ken Malmberg and Andrew Heathcote for comments on an earlier draft. This work was supported in part by a grant from the US Air Force Office of Scientific Research (AFOSR grant number FA9550-04-1-0317) to Steyvers & Brown, "Inference in Dynamic Environments".

---

[1] Starting values for the searches were calculated by estimating static SDT models separately for easy and hard decision contexts. Independent minimizations were carried out for all feasible values of the lag parameter. "Feasible" lag values included any positive integer not greater than the block length in Experiments 2-4. In Experiment 1, where the stimulus switch point was within a block, feasible lags included some negative values, allowing that some participants may have anticipated the stimulus switches.