Journal of Experimental Psychology:
Learning, Memory, and Cognition
2005, Vol. 31, No. 4, 587–599

0278-7393/05/$12.00 DOI: 10.1037/0278-7393.31.4.587

The Dynamics of Experimentally Induced Criterion Shifts

Scott Brown and Mark Steyvers
University of California, Irvine

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 participants' decision criteria. When easy and difficult conditions are blocked alternately and a mirror effect is observed, participants must repeatedly change their decision criteria. The authors investigated the time course of these criterion changes and observed the buildup of mirror effects on a trial-by-trial basis. The 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 has implications for traditional blockwise analyses of data and for models of decision making.

Keywords: mirror effect, dynamics, context effect, signal detection theory, decision making

A common assumption in models of decision making is *stationarity*. With few exceptions (e.g., Kac, 1966; Rabbit, 1981; Strayer & Kramer, 1994a, 1994b; Treisman & Williams, 1984; Vickers & Lee, 1998, 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 used stationary decision-making environments. More recently, there has been a growing focus on nonstationary (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 and error-rate tracking (e.g., Heit, Brockdorff, & Lamberts, 2003; Kelly, Heath, & Longstaff, 2001; Petrov & Anderson, 2005; Rotello & Heit, 2000; Treisman & 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 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 perhaps 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. Below, 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. For generality, we refer to the two classes of decision stimuli as *targets* and *distractors*. These stimuli could be malignant versus benign tumors, 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 environment, the distractors and targets 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.

The alternating decision contexts illustrated 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

Scott Brown and Mark Steyvers, Department of Cognitive Sciences, University of California, Irvine.

We thank Eric-Jan Wagenmakers, Ken Malmberg, and Andrew Heathcote for comments on a draft of this article. This work was supported in part by Grant FA9550-04–1-0317 from the U.S. Air Force Office of Scientific Research to Mark Steyvers and Scott Brown.

Correspondence concerning this article should be addressed to Scott Brown, Department of Cognitive Sciences, 3151 Social Sciences Plaza, University of California, Irvine, CA 92697-5100. E-mail: scottb@uci.edu

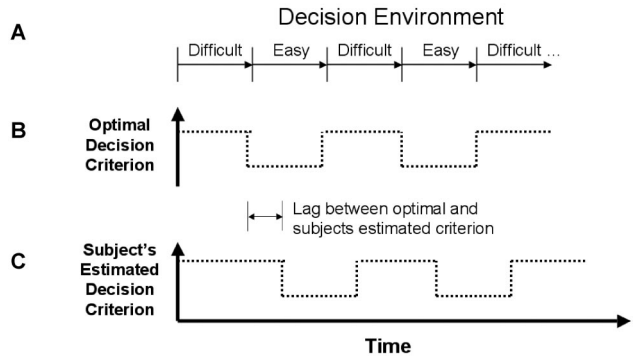

Figure 1. Basic paradigm. Decision environments change (A) leading to changes in the optimal decision criterion (B). Participants' actual decision-making processes (C) lag behind.

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 (e.g., distractor items) are changed. Changes in FAR are to be expected from changes to the stimuli with which they are associated (distractors) but mirror effects include changes in HR that cannot be explained this way. Because 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 through the use of SDT (Green & Swets, 1966). SDT posits that participants decide between two classes of items (i.e., 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 to 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 Figure 2A 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 our analyses we always allow arbitrary criterion values (i.e., we estimate bias).

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 Figure 2B), 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 resulting from 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. Although 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 is 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

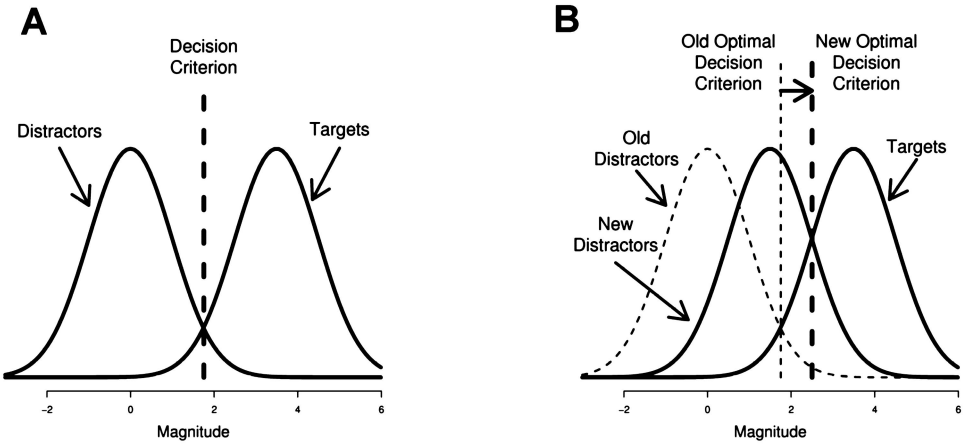

Figure 2. A: Illustrates standard signal detection theory. B: How the optimal decision criterion changes when the properties of the distractors are altered. Stimuli above the decision criterion are classified as targets; stimuli below are classified as distractors.

the difficult environment, defined by a sensitivity parameter (d'_H, where H = hard) and a decision criterion (C_H) and another SDT model operating in the easy decision environment, defined by d'_E and C_E (where E = easy). We assume equal variances for the target and distractor distributions, although later we discuss—and test—unequal variance models.

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 the stimuli themselves define decision difficulty. By contrast, the decision criterion is under the control of the decision makers, 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. Our assumption of a stepwise 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 trials after an environment change before participants change their criterion).

Model Predictions

Figure 3 illustrates the predictions of this model. Once again, we have drawn optimal decision criteria on Figure 3 for simplicity of illustration; when fitting the model to data, we allow for nonoptimal criteria by estimating bias parameters. The SDT model depicted in Figure 3A, 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 HR and FAR predictions at the left edge of Figure 3F, 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 Figure 3B, labeled t_2: note that sensitivity has decreased as a result of the harder stimuli (d'_E has changed to d'_H), but the criterion has not yet changed. This leads to

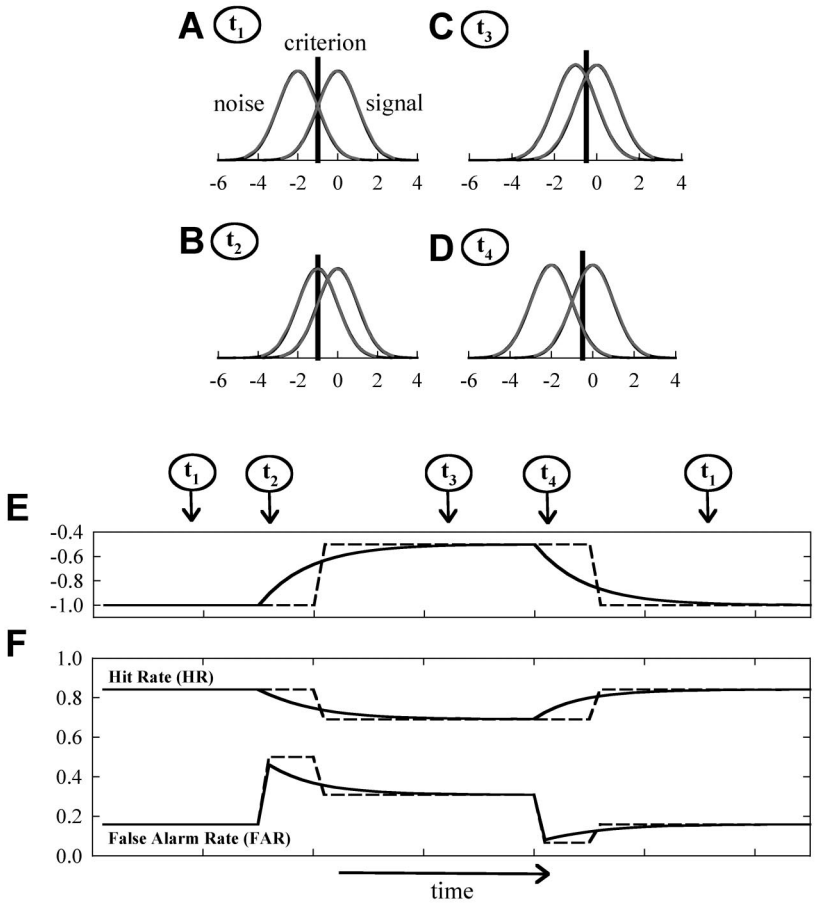

Figure 3. Predictions from the dynamic signal detection theory (SDT) model. A–D: Static SDT submodels. E: Predicted criterion changes. F: Predicted hit and false-alarm rate changes.

the HR and FAR predictions shown in Figure 3F in dashed lines under t_2: no immediate change in HR but a large increase in FAR. After some lag, L, the decision maker updates his or her criterion to C_H, the criterion for hard decision environments (shown by the dashed line in Figure 3F). The SDT model then operating is shown as t_3 in Figure 3C, and its predictions are shown by the dashed lines under the label t_3 in Figure 3F: 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 shown in Figure 3D 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 3E and 3F show predictions for an individual participant: 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 participants for graphing. Those graphs show smooth changes in FAR and HR, as illustrated by solid lines in Figure 3E and 3F. Smooth transitions are the result of averaging over many individual stepwise transitions with variable step positions.

Equal Versus Unequal Variance Assumptions

The illustrations of the dynamic SDT model above all use equal variance target and distractor distributions. In recognition memory tasks, as opposed to the lexical decision task we use, many researchers have observed unequal variances for these distributions (e.g., Glanzer, Kim, Hilford, & Adams, 1999; Malmberg, 2002; Ratcliff, Sheu, & Gronlund, 1992; Sheu & Heathcote, 2002; Verde & Rotello, 2003). We have no a priori reason to believe that the target and distractor distributions should have unequal variances in lexical decision. On the contrary, there is some evidence for an equal variance assumption. Jacobs, Graf, and Kinder (2003) measured receiver operating curves in a lexical decision task and observed that the slope of their z-transforms was not significantly different from one, supporting the equal variance assumption. Our model analyses are not restricted to the equal variance assumption, so we provide direct tests of whether unequal variance models provide better descriptions of the data. Additionally, we test whether the relevant parameter estimates (criterion lags) are different in equal versus unequal variance models. To foreshadow, the unequal variance models do not significantly improve goodness of fit, nor do they significantly alter parameter estimates.

Change Mechanism

A limitation of the model so far is that it does not include a mechanism for how changes in criterion occur: What triggers a change? Mechanisms have been proposed on the basis of response monitoring (Treisman & Williams, 1984), stimulus monitoring (Strayer & Kramer, 1994b; Vickers & Lee, 1998, 2000), and error-rate monitoring (Rabbit, 1981). Each of these mechanisms entails constant, small-scale adjustment to criterion position. We use a simpler, descriptive model of criterion change that ignores small, spontaneous changes and returns focus to the larger, experimentally induced changes. This mechanism is based on change detection and so applies naturally to experimentally induced criterion changes, which presumably depend on detection of changed stimulus properties.

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 for 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 in which participants classify strings of letters as words (e.g., *CAT*) or nonwords (e.g., *CXT*; see, e.g., Glanzer & Ehrenreich, 1979; Gordon, 1983; Grainger & Jacobs, 1996; Ratcliff, Gomez, & McKoon, 2004).

Wagenmakers et al. (2004) 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 the Kello and Plaut (2000, 2003) tempo naming task. This procedure allows us to hold RT relatively constant and to observe changes only in response accuracy, which simplifies analysis and allows comparison with the predictions of our dynamic SDT theory.

We used a lexical decision task in 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 block design in which the decision context only 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, participants decided whether strings of arrows had more arrows facing left or right. The properties of one kind of display (left or right facing) were kept constant within participants, whereas the properties of the other were manipulated to change decision difficulty. Changes in difficulty occurred only during block breaks.

Method

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 it was not (mouse button timing on our systems provides an accuracy of about ±6 ms; see Beringer, 1992). The buttons used for each response were counterbalanced across participants. The same set of words and nonword strings were used in all three experiments. Seven-letter words were drawn from the Kučera and Francis (1967) word pool, and nonwords were constructed by altering either just one letter of a valid word (making *hard* nonwords) or by altering three letters (*easy* nonwords), always checking to ensure 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.

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 the Kello and Plaut (2000, 2003) tempo naming task. In their procedure, a series of rhythmic tones are presented on each trial that help participants anticipate the response signal. We generalized their method by keeping a constant rhythmic tone throughout each experimental block, continuing between trials. Every 400 ms, a 256-Hz tone sounded for 50 ms, 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 330 ms and 700 ms 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 participants 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 informed of 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., interbeep time of 600 ms), 33% in the second block (interbeep time of 533 ms), and 16.7% in the third block (interbeep time of 467 ms). 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 7 ms in any timing setting, and stimulus presentations were synchronized with the screen's vertical refresh.

Table 1
Examples of Easy and Hard Nonword Strings and Words Used in Experiments 1–3

Nonwords		
Easy	*Hard*	Words
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*

Note. The "wordiness" of easy nonwords is lower than that of the hard nonwords.

Experiment 1: Details. Participants were 149 undergraduates from the University of California, Irvine, who received course credit for participating. Data from participants 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 or that there were different classes of stimuli.

Experiment 2: Details. Participants were 108 undergraduates from the University of California, Irvine, who received course credit for participating. Data from only 2 participants (1.85%) were rejected as a result of 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 5 words or nonwords in succession. As before, the words used were always drawn from the same pool throughout the experiment, whereas the nonwords alternated from easy to hard across blocks. The class of nonwords used for the first block (easy vs. hard) was randomized across participants. Participants were not informed about the classes of stimuli.

Experiment 3: Details. Participants were 169 undergraduates from the University of California, Irvine, who received course credit for participating. Data from 7 participants (4.1%) were rejected as a result of poor performance. The design for Experiment 3 was identical to that of 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. hard) 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, although we used 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 to the right and to push the corresponding mouse button. The left-to-right order of the arrows was randomly shuffled on every trial.

We used the same rhythmic beeping procedure 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

<<<><<<<<<

><><<<>><<

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.

distribution of proportions of left and right facing elements used: a display with 4 to 6 left–right elements (like the lower stimulus in Figure 4) is much more difficult than one with 1 to 9 left–right elements (like the upper stimulus in Figure 4). We always had right-favoring displays with either 4 to 6 or 3 to 7 proportions. We varied (between blocks) the left-facing displays from easy (either 9 to 1 or 8 to 2) to hard (either 7 to 3 or 6 to 4). We reversed the left–right assignment for half of the participants, although we collapsed 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 HR and FAR for our participants, separately for blocks with easy and hard decision contexts. Figure 5 shows these data averaged across participants for Experiment 1.

The data from Experiment 1 are shown in Figure 5A, aligned at the switch point from easy-to-hard (solid symbols) or from hard-to-easy (open symbols) decision contexts. The smooth lines are predicted probabilities from the dynamic SDT model—the HRs show smooth changes even though our model assumes stepwise criterion changes simply because of averaging across participants in the plot. The data were also averaged over blocks of eight consecutive trials for the purposes of graphing. Note that model fitting was done on completely unaveraged (trial-by-trial) data. The *y*-axis shows the probability of "word" responses. The FARs show that our manipulation of decision difficulty had the desired effect. When the decision context was easy (solid symbols), the probability of incorrectly identifying a nonword as a word was low (FAR, which is represented by 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 hard to easy (open symbols).

More interestingly, the change in nonword properties resulted in changes in the responses to word stimuli, shown by the HR (represented in circles in Figure 5). After the nonwords changed from easy to hard (solid symbols), the HR steadily declined; 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 reemerged. 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 Figure 5B, which plots the difference in HRs 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 8-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 switch, $t[134] = 2.1$; 25–32 trials after switch, $t[134] = 3.3$; all ps $< .05$). Note once again, that the smooth changes observed in HR and FAR are consistent with averaging stepwise changes across participants.

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 toward 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.

Experiments 2 and 3

Data from Experiments 2–4 are shown in Figure 6 in the same format as Figure 5. The data from Experiments 2 and 3 show

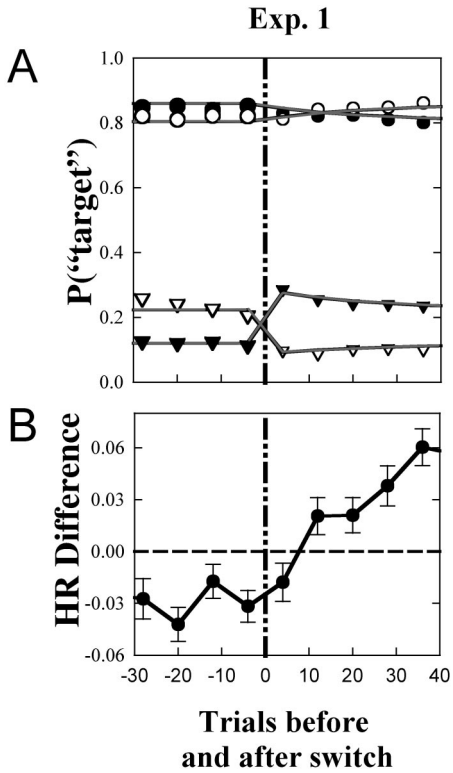

Figure 5. Data from Experiment 1 aligned to the stimulus switch point. Solid symbols correspond to blocks in which decision contexts changed from easy to hard, and open symbols represent blocks in which decision contexts changed from hard to easy. A: Circles represent hit rate (HR); triangles represent false-alarm rate. B: The difference between HR in easy and hard conditions changing within a block. P = probability.

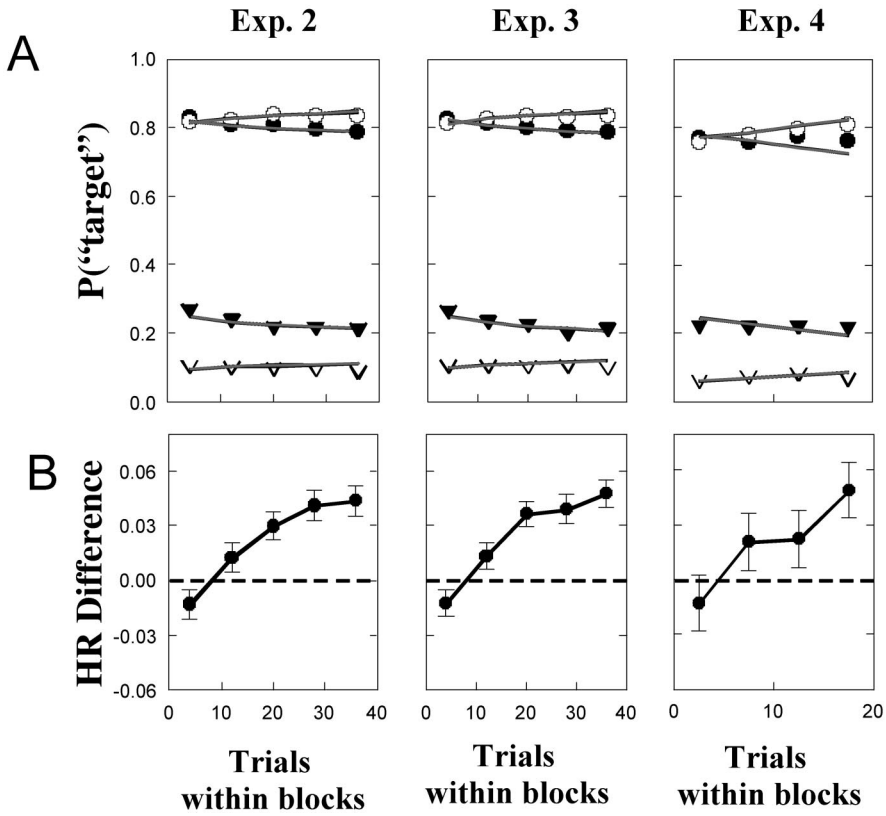

Figure 6. Data from Experiments 2–4. The *x*-axes show trials within each block. Solid symbols correspond to hard decision contexts, and open symbols correspond to easy decision environments. A: Circles represent hit rate (HR); triangles represent false-alarm rates B: The difference between HR in the easy and hard conditions. P = probability.

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 0, 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 (solid 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., Trials 33–40 in Experiment 2, $t[105] = 5.2$, $p < .001$; Trials 33–40 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). HRs in the easy condition were smaller than in the hard condition during the first 8 trials of each block (the mean difference was 1.3%, $t[105] = 1.6$, $p = .055$, in Experiment 2; the mean difference was 1.2%, $t[161] = 1.7$, $p < .05$, in Experiment 3). In Experiment 3 a statistically significant mirror effect reemerged 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 FARs from Experiment 3 demonstrate the same slow changes as observed in Experiment 1 (although possibly to a smaller extent), and this is consistent with our lagged criterion change explanation of these data. After the stimulus properties changed (between blocks), there was a sudden large change in FARs. The FARs then slowly changed over the course of the block toward more moderate values. The changes in FARs from Experiment 2 were as expected (i.e., in the same direction as HR changes) for difficult blocks only. For the easy blocks there was a very slight trend for the FARs 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 close to statistical significance (simple effects analysis of variance: $F[4, 101] < 1$), but they are in the opposite direction to our predictions. Our assumption of nonoptimal criterion placement can change predictions about slow changes in FAR.[1] If the criterion placement in hard decision blocks is close to optimal, and the criterion placement in easy decision blocks is higher than optimal, our model would predict very small (or no) smooth changes in FARs when changing from hard to easy contexts, though still predicting nontrivial trends in FARs for easy-to-hard context changes (as observed). Indeed, these asymmetric biases seem to be present in our data. Our estimates of decision criterion placement below show that, for each experiment, participants reliably dem-

[1] We thank Vincent Stretch for pointing this out.

onstrated greater bias (toward target responses) in easy than in hard blocks: Over all experiments, the mean bias (criterion estimate standardized by d' estimate) was 24% larger in easy than in hard decision contexts. These differences were reliable in each experiment: Experiment 1, $t(134) = 8.0$, $p < .001$; Experiment 2, $t(105) = 3.1$, $p < .01$; Experiment 3, $t(161) = 2.8$, $p < .01$; and Experiment 4, $t(128) = 2.0$, $p < .05$. These observed bias differences support the above explanation of small or zero FAR changes after switching.

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 were unable to adjust to new decision environments more quickly even when recruiting conscious, intentional processes. Strayer and Kramer (1994a, 1994b) found similar effects in their experiments: Participants were unable to speed up adjustments of their (speed–accuracy trade-off) criteria in response to instructions.

Experiment 4

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 5 trials, rather than the previous 8, because of the smaller block lengths. The data from Experiment 4 were very similar to those from Experiments 2 and 3. When the decision task was easy, the FARs were 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 5 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 HRs were 2.1% higher than were hard HRs, 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 FARs and higher HRs. 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 reinstate a mirror effect. Similar slow changes in FARs were observed, that were always in the same direction as changes in HRs. 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 biases 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 stepwise 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 unaveraged data, it is simple to calculate maximum likelihood estimators of the parameters by search.[2] The predicted HRs and FARs for this model are shown by the solid lines in Figures 5A (for Experiment 1) and Figure 6A (Experiments 2–4), aggregated in the same way 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.

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 0 and 5 trials. Some participants behaved very differently: 8% had negative estimated lags, and 20% had estimated lags greater than 40 trials. Averaged over participants, 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, which again suggests that knowledge of the experimental design and specific instructions to participants do not decrease the amount of time taken to adjust to new decision environments (consistent with Strayer & Kramer, 1994a, 1994b). In Experiments 2 and 3, averaged over participants, 13 and 15 trials, respectively, were required for participants 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 investigated unequal variance models for our data. Our methodology allows us to estimate variance parameters with standard maximum-likelihood techniques, simultaneous with estimation of the other

[2] 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.

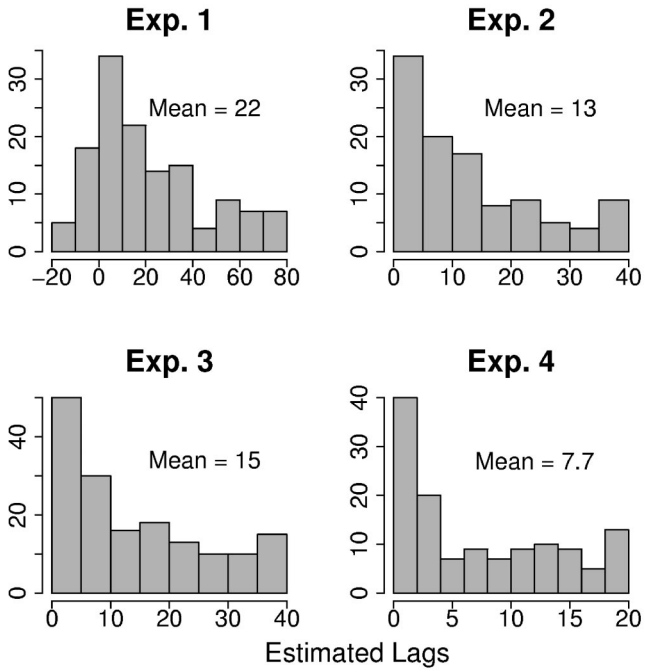

Figure 7. Histograms of estimated lags from fits of the dynamic signal-detection theory model. Exp. = experiment.

model parameters.[3] 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, thus making 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 chi-square 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. We fit a model 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 than 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 were little data for this estimation, so numerical difficulties could have resulted in nonoptimal fits. As a check against this, we fixed the variance parameters across all participants 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 (Akaike's information criterion, the Bayesian information criterion, 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 one half. 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, $t(161) < 1$.

Practice Effects

Given the length of the experiments, it is natural to wonder whether there were practice effects. Perhaps the lag in criterion change decreased in the latter parts of the experiments, as subjects became more adept at anticipating change; or perhaps the lag increased in the latter parts of the experiments, due to fatigue. We tested these effects by separately fitting dynamic SDT models to the first and second halves[4] of each participant's data. We then calculated repeated-measures t tests to assess whether the criterion lag estimates differed in these models, and whether the magnitude of the criterion change differed (i.e., did the size of the mirror effect change?). The results are presented in Table 2. In Experiment 1, participants demonstrated longer criterion lag times and smaller criterion shifts in the second half of the experiment than in the first half. These changes may be due to fatigue effects: Experiment 1 was longer and more arduous than were the other experiments, which was due to longer blocks and greater number of trials. There was also a smaller but statistically significant increase in the magnitude of the criterion shift in the second half of Experiment 4 compared with the first half.

Summary of Results

The experiments and data analyses presented above demonstrate the dynamic build-up of mirror effects over time. Mirror effects

[3] The assumption of a lag parameter allows this estimation. Lags result in measurements of hit and false-alarm rates for the same stimulus classes with different response criteria, just as with confidence rating data.

[4] Experiment 1 had only 10 blocks, so we separately analyzed Blocks 1–5 and 6–10. This ensured an equal number of hard-to-easy and easy-to-hard transitions.

Table 2
Results of Repeated Measures t Tests for Experiments 1–4

		Lag difference			Criterion change size		
Experiment	*df*	*M* difference	*t*	*p*	*M* difference	*t*	*p*
1	134	−5.00	1.70	.044*	0.099	5.10	<.01*
2	105	0.75	0.49	>.10	0.026	1.20	>.10
3	161	−2.00	1.50	>.05	0.020	1.20	>.10
4	128	−0.39	0.43	>.10	−0.045	1.70	.044*

Note. Mean differences are defined as means of lags or criterion change sizes in the data from the first half of each experiment minus those means from the second half of each experiment. Results indicate differences in criterion change size and lag estimates in the first versus the second half of data.
* Significant at the $p = .05$ level.

were established by changing the difficulty of decisions, and HR and FAR changes were observed following changes in decision difficulty. We developed a simple dynamic version of SDT in which the decision criterion changes some time (L) after stimulus properties change, and we used this model to fit data at an individual-participant 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 (and, respectively, easy) decision contexts actually reflects participants' easy (and, respectively, 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 blockwise 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 blockwise 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 sample trials of each block, choosing sample size to maximize the observed effect size. This involved a trade-off between increasing HR differences and increasing variability as a result of decreasing sample sizes. For Experiment 2, the maximum effect size was 0.70 standard deviations (an increase of 11%), which occurred 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 of 10%), which occurred when we removed the first five trials of each block.

A Change Detection Model

In Experiments 1–4, participants were required to make a decision about every stimulus. Optimum performance in the lexical decision or numerosity categorization task required participants to detect and respond to changes in their stimulus environment. The change detection task was thus secondary to the central decision task. This contrasts with other statistical research investigating change detection. Such research usually begins with the aim of detecting a "change point" in a complete sequence of data. In our paradigm, participants had to detect changes in a sequence as that sequence unfolded. It is natural to ask not only how well do participants perform in the main task, but also what mechanism allows them to detect changes. Though many mechanisms are possible and are consistent with the data, we investigate a simple model based on null hypothesis significance testing.

An example that illustrates the model is one in which an observer is presented with a sequence of numbers and he or she is told that at exactly one point during the sequence the distribution of these numbers will change. The observer's task is to identify when the change has occurred, as the sequence unfolds. After each data point is presented, the observer is asked "Has the switch point passed yet?", and he or she is allowed to answer "yes" only one time. This situation parallels that in our experiments in which there was one context change per block. Our change detection model assumes a nested model comparison framework for this situation. Suppose that the pre- and postswitch data are drawn from normal distributions (as per the assumptions of the dynamic SDT model). Given parameter estimates for those distributions, it is simple to evaluate the evidence for two nested models. The first model is that the data so far have been generated by just one distribution: The likelihood of the data under this hypothesis is easy to compute. The second hypothesis is that the data so far were generated from the one distribution up to some hypothetical switch point and then from another distribution after that point. The likelihood of this hypothesis could be calculated by evaluating the likelihood separately for each possible switch point (from the beginning of the sequence to the current time) and choosing the most likely switch point. Continuing the spirit of maximum likelihood estimation, we assume that the observer estimates the means of the distributions directly from data.

The model would then be in a position to respond that the switch had occurred if the likelihood of the switch model exceeded the likelihood of the one-state model by some criterion level. Under the above assumptions, twice the difference in log-likelihood values for the two nested models will be distributed as a chi-square variable with two degrees of freedom (one for the extra distribution mean parameter, one for the switch point location parameter). The criterion amount of difference required to detect a change would be determined by the ideal observer's desired Type I error rate. If the ideal observer has a very strict Type I error rate (e.g., $p = .0001$), he or she will only decide that a change has occurred when the two-state model has very much greater likelihood than the one-state model (larger by about 18.4 units). Conversely, if the ideal observer has a very lenient Type I error rate (e.g., $p = .10$)

he or she will decide that change has occurred when the two-state model fits only a little better than the one state model (larger likelihood by only 4.6 units). Adjustments in the Type I error rate implement a trade-off between the confidence that a change actually has occurred when a response is made and the lateness of that response.

We implemented the ideal observer model individually for each participant as follows. Measures of decision sensitivity in both easy and hard decision environments were obtained from our dynamic SDT measurement model (d'_E and d'_H). The difference in these d' values represents the stimulus change that participants must detect during our experiments. For each participant in Experiments 2–4, we simulated 1,000 random sequences of the same length as the experimental sequences. These sequences alternated between standard normal data, $N(0, 1)$, and $N(d'_E - d'_H, 1)$, with alternations occurring at the same frequency as in the experiments (i.e., each block). Experiment 1 was not used, as negative lags were possible in that experiment. Also, data from participants with $d'_E - d'_H < 0.75$ were not used for these analyses, as small stimulus property changes result in large estimation error for the detection model. With the simulated sequences, and for any given level of evidence required for change detection (Type I error rate or p value), the distribution of change detection lags for the detection model was estimated by Monte Carlo integration. For each participant, we identified the Type I error rate that provided the maximum likelihood fit of the distribution of change detection lags from the model to the change detection lag estimated earlier from our dynamic SDT model. That is, we found the Type I error rate that resulted in a distribution of change detection values with mode closest to the estimated lag from the SDT model.

Figure 8 shows estimates of the Type I error rate (p value) plotted against estimates of the lag parameter from the dynamic SDT model for Experiments 2–4. There are strong negative correlations evident, as shown by the R^2 values and best-fitting lines. These correlations show that fits of the change detection model are capturing the notion that short lags are associated with relatively large (i.e., lenient) Type I error rates and vice versa. The correlations are less than 1 because different participants have different d' changes to detect (i.e., the subjective magnitude of the difficulty manipulation varied across participants). Combining the dynamic SDT model fits with the change detection model provides a process model for change detection and performance in our paradigm.

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 criterion shifts should be observed and what those shifts should look like, which is a departure 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: block 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 researchers do use dynamic analyses, but mostly they use 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 describes only the time course of criterion shifts, without addressing the question of the causation of the shifts. There are a multitude of plausible theories to explain how

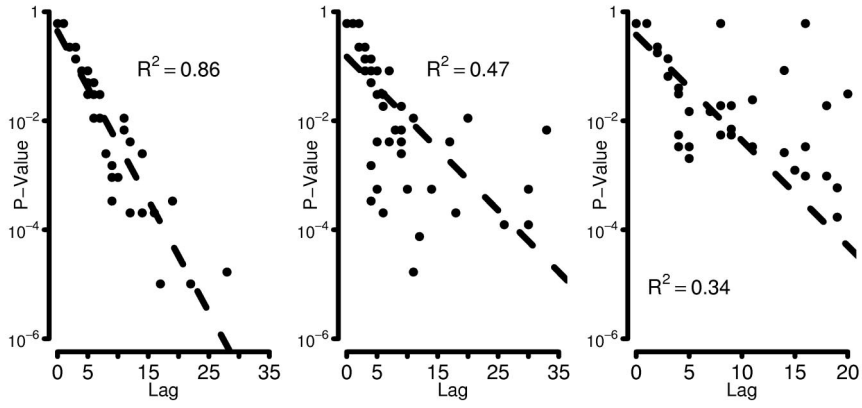

Figure 8. Data from Experiments 2–4 (left to right). Plots of Type I error rate for the change detection model (p value, y-axis) against lag estimates from the dynamical signal detection theory model (x-axis) Also plotted are best-fitting lines in log-linear space, with corresponding R^2 values.

participants adjust their decision criteria. The dynamic version of SDT we use is similar to the Treisman and Williams (1984) theory for criterion setting. In their theory, participants are assumed to change their decision criterion after each trial on the basis of the stimuli and responses from the last few trials. Our dynamic SDT allows for a lag parameter that measures how far behind the stimulus change participants change their decision criteria and is thus a simplified version of the Treisman and Williams model. The value of Treisman and Williams's 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 (i.e., *response monitoring*). For example, the obvious extension of Treisman and Williams's 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 to 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 stimulus 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. distinguish these effects by 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, hard 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 when they are 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. Although this method of modeling the data was 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 as 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 by using one or other of two decision "modules"—one module that is best suited for use in easy decision contexts and one that is best suited for use in hard decision 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).

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 9. 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 nonlagged 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 used 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 hard-decision environments, so that behavior in each environment is influenced by the previous environment for many trials. We show that these effects are both quali-

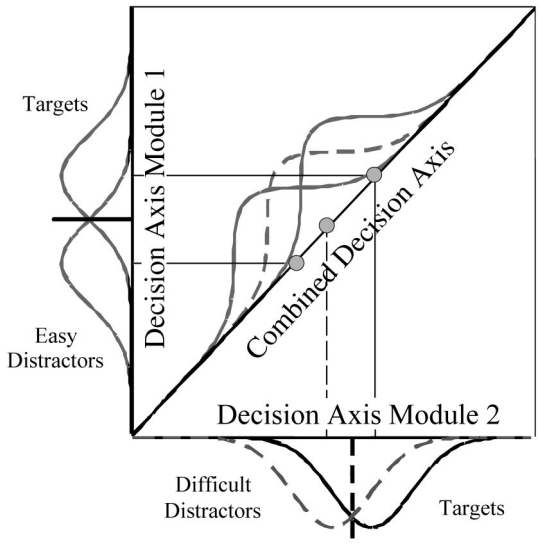

Figure 9. Isomorphism (at a descriptive level) of simple strategy shift and criterion shift models.

tatively 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, & Computers, 24,* 486–490.

Gilden, D. L., Thornton, T., & Mallon, M. W. (1995, March 24). 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,* 8–20.

Glanzer, M., & Adams, J. K. (1990). The mirror effect in recognition memory: Data and theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 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, and Cognition, 25,* 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,* 518–565.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley

Heit, E., Brockdorff, N., & Lamberts, K. (2003). Adaptive changes of response criterion in recognition memory. *Psychonomic Bulletin & Review, 10,* 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,* 481–488.

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

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,* 719–750.

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.

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, 55*(A), 805–840.

Kučera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English.* Providence, RI: Brown University Press.

Malmberg, K. J. (2002). On the form of ROCs constructed from confidence ratings. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 380–387.

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. (2005). The dynamics of scaling: A memory-based anchor model of category rating and absolute identification. *Psychological Review, 112,* 383–416.

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,* 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,* 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,* 1379–1396.

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

Van Orden, G. C., Holden, J. G., & Turvey, M. T. (2003). Self-organization of cognitive performance. *Journal of Experimental Psychology: General, 132,* 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,* 739–746.

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,* 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,* 1–31.

Wagenmakers, E. J. M., Steyvers, M., Raaijmakers, J. G. W., Shiffrin, R. M., van Rijn, H., & Zeelenberg, R. (2004). A model for evidence accumulation in the lexical decision task. *Cognitive Psychology, 48,* 332–367.

Received May 4, 2004
Revision received October 26, 2004
Accepted December 27, 2004 ■