Journal of Personality and Social Psychology
1991, Vol. 61, No. 3, 380-391

0022-3514/91/$3.00

Automatic Vigilance: The Attention-Grabbing Power of Negative Social Information

Felicia Pratto and Oliver P. John
University of California at Berkeley

One of the functions of automatic stimulus evaluation is to direct attention toward events that may have undesirable consequences for the perceiver's well-being. To test whether attentional resources are automatically directed away from an attended task to undesirable stimuli, Ss named the colors in which desirable and undesirable traits (e.g., honest, sadistic) appeared. Across 3 experiments, color-naming latencies were consistently longer for undesirable traits but did not differ within the desirable and undesirable categories. In Experiment 2, Ss also showed more incidental learning for undesirable traits, as predicted by the automatic vigilance (but not a perceptual defense) hypothesis. In Experiment 3, a diagnosticity (or base-rate) explanation of the vigilance effect was ruled out. The implications for deliberate processing in person perception and stereotyping are discussed.

There is a fundamental asymmetry in people's evaluations of gains and losses, of joy and pain, and of positive and negative events. A considerable body of research, in fields as diverse as decision making, impression formation, and emotional communication, has shown that people exhibit loss aversion (Kahneman & Tversky, 1984): They assign relatively more value, importance, and weight to events that have negative, rather than positive, implications for them. In decision making, potential costs are more influential than potential gains (e.g., Kahneman & Tversky, 1979). In impression formation, negative information is weighted more heavily than positive information (e.g., Anderson, 1974; Fiske, 1980; Hamilton & Zanna, 1972). In nonverbal communication, perceivers are more responsive to negatively toned messages than to positive ones (Frodi, Lamb, Leavitt, & Donovan, 1978). Quite generally, then, "losses loom larger than gains" (Kahneman & Tversky, 1984, p. 348).

There are good evolutionary reasons for this widespread and pronounced asymmetry in people's evaluative reactions. Events that may negatively affect the individual are typically of greater time urgency than are events that lead to desirable consequences. Averting danger to one's well-being, such as preventing loss of life or limb, often requires an immediate response. In comparison, positively valenced activities, such as feeding and procreation, are less pressing; although they are of crucial importance in the long term, pleasure is simply less urgent than pain. Negative affect carries an important signal value because it signifies to the organism the need to change or adjust its current state or activity.

Given the adaptive significance of fast responses to undesirable stimuli (e.g., Fiske, 1980), an adaptive advantage would accrue for organisms that have the capacity to attend to them quickly and with little effort. In humans, quick and effortless cognitive processes have been termed *automatic;* that is, they can occur without the perceiver's intention or control (for a review, see Shiffrin, 1988). In this article, we postulate and provide evidence for *automatic vigilance,* a mechanism that serves to direct attentional capacity to undesirable stimuli. Previous research suggests that people automatically process the subjective evaluation of social stimuli, such as liked and disliked attitude objects (Fazio, Sanbonmatsu, Powell, & Kardes, 1986), and this research is reviewed below. The present studies of automatic vigilance build on these earlier demonstrations of automatic evaluation effects and are designed to show that undesirable social stimuli are more likely to attract attention than are desirable social stimuli.

Automatic-Evaluative Processing

Evaluation of stimuli as good or bad, liked or disliked, and desirable or undesirable is a basic and ubiquitous aspect of the way people respond to their environment in both its social and nonsocial aspects. A variety of psychological theories view evaluation as a central and even primary response. For example, in factor analyses of semantic differential ratings, Osgood, Suci, and Tannenbaum (1957) found that evaluation was the first and largest factor of connotative meaning, a finding replicated across numerous cultures and languages. Lazarus (e.g., 1966, 1982) has suggested that emotions depend on the person's appraisal of an event and that one aspect of primary appraisal is the simple, immediate ascertainment of whether a stimulus is "good for me" or "bad for me."

This research was supported, in part, by National Institute of Health Grant MH39077 and by Biomedical Research Grants 87-20 and 88-24 (University of California) to Oliver P. John.

We are indebted to Celeste Schneider and Kristina Whitney Robins for assisting as experimenters, to Bud Viera for his help in decoding the MEL programming language, and to John Bargh, Asher Cohen, Lewis R. Goldberg, Shinobu Kitayama, Richard Lazarus, Delroy Paulhus, Steven J. Sherman, Arthur Shimamura, and Shelley E. Taylor for their comments on a draft. The support and resources of the Institute of Personality Assessment and Research, where Felicia Pratto spent 2 postdoctoral years, are also gratefully acknowledged.

Correspondence concerning this article should be addressed to Felicia Pratto, who is now at the Department of Psychology, Jordan Hall, Building 420, Stanford University, Stanford, California 94305-2130, or to Oliver P. John, Department of Psychology, University of California, Berkeley, California 94720.

Consistent with this view of evaluation as primary appraisal, a number of recent studies suggest that people can and do evaluate stimuli easily, readily, and quickly and that sometimes they do so without intention or much conscious thought. For example, when subjects who had no particular processing goal were exposed very briefly (1–3 ms) to physical shapes, their liking judgments of previously presented shapes differed from those of new shapes (Kunst-Wilson & Zajonc, 1980; Seamon, Brody, & Kauff, 1983; Seamon, Marsh, & Brody, 1984). The accuracy of affective choices was found to exceed recognition accuracy at short exposures (Seamon et al., 1983, 1984), although subjects became accurate at recognition judgments at only slightly longer exposures (Seamon et al., 1984).

Accurate evaluative judgments can also be made for semantically meaningful material at exposure speeds at which accurate recognition does not occur. For example, Bargh, Litt, Pratto, and Spielman (1988) presented trait adjectives at speeds below each subject's threshold of stimulus recognition (i.e., subjects could not accurately recognize whether a stimulus word or a blank card had been presented). Subjects were able to make correct evaluative judgments although they were unable to make correct synonymy judgments. Thus, subjects were able to ascertain the evaluative information in symbolic representations (words) before recognition.

In research on attitudes, Fazio (e.g., 1986; Fazio et al., 1986) found that the evaluation associated with a person's attitude toward an object becomes accessible automatically on exposure to relevant stimuli and that attitude–behavior consistency can be explained in terms of this accessibility. The role of evaluation in intergroup perception has been addressed by Fiske (e.g., 1982), who argued that evaluation becomes immediately accessible when stereotypes are activated.

In all, this research suggests that the evaluation of a stimulus can be detected before conscious recognition occurs and that evaluation is one of the first aspects of semantic meaning to be ascertained. Although it appears to be widely assumed that the evaluative aspects of many different types of stimuli are processed quickly and without much effort, differences between the desirable and undesirable poles of the evaluative continuum have not been examined systematically. The present studies were designed to demonstrate experimentally that automatic attention to desirable and undesirable stimuli is asymmetrical.

Testing the Automatic Vigilance Hypothesis

One way to demonstrate that an automatic mechanism directs attention to undesirable stimuli is to use a task in which the evaluative component of the stimulus is irrelevant to task performance but may interfere with it. Interference with the performance of an attended task is usually taken as an indicator of automaticity: "If a process produces interference with attentive processes despite the subject's attempts to eliminate the interference, then the process in question is surely automatic" (Shiffrin, 1988, p. 765). The key feature of an automatic process is thus its inescapability.

In the present studies, we used a task modeled after Stroop's (1935) color-interference paradigm. In the standard Stroop task, subjects name the colors in which a set of words is presented; attending to the meanings of the words leads to interference, particularly when the letters spell a color name different from the color in which the word is printed. To test whether attention is directed to negatively evaluated stimuli even when subjects try not to attend to that aspect of the stimuli, we presented a series of personality-trait adjectives, such as *sadistic, honest,* and *outgoing,* and subjects named the color in which the adjectives were presented. We predicted that although they had no intention or reason to do so, subjects would attend more to undesirable than to desirable traits and that this additional attention would lead to relatively longer color-naming latencies for undesirable traits. In our stimulus sets, we included traits covering a wide range of evaluation, ranging from extremely undesirable (e.g., sadistic, wicked, mean) to extremely desirable (e.g., kind, friendly, honest). This allowed us to examine the relation between the desirability values of the traits and subjects' latencies in naming their colors.

When rating the desirability of a set of traits, subjects can make fine-grained distinctions among traits on the evaluative continuum. One might therefore assume that gradations in trait desirability would affect color-naming latencies. In that case, trait desirability would be linearly related to color-naming latency, as shown in the middle panel of Figure 1. Across the whole range of desirability values, the more negative the trait, the more it should attract the perceiver's attention and distract from the color-naming task. In correlational terms, one would expect a negative association between desirability and latency even within the two valence categories.

However, we have argued that one function of automatic evaluation is to monitor the environment for undesirable stimuli. If this hypothesis is correct, it may be of little importance exactly how undesirable the stimulus is. It would be sufficient for the initial screening to tag any potentially undesirable event; the specific meaning of the tagged stimulus, including its severity, can be ascertained by subsequent, more controlled, processing. Moreover, automatic evaluations seem to occur very rapidly and early in processing, so that the evaluative distinctions afforded by this process may be relatively crude, possibly no more complex than a simple categorical distinction between desirable and undesirable (see also Bargh et al., 1988). The relation between trait desirability and color-naming latencies might then take the categorical form depicted in the top panel of Figure 1. According to that hypothesis, desirable and undesirable traits differ from each other in their color-naming latencies; within the two valence categories, however, desirability and latency should not be related.

We also examined whether evaluative extremity might influence color-naming latencies. If that were the case, more extremely undesirable and desirable traits should elicit the longest latencies, with latencies decreasing from the extremes towards the more neutral traits, as shown in the bottom panel of Figure 1. In summary, we examined three different types of relations between desirability and response latency: categorical, linear, and quadratic.

Experiment 1

Method

Subjects. Sixteen undergraduates at the University of California at Berkeley participated in exchange for partial course credit; data from 5

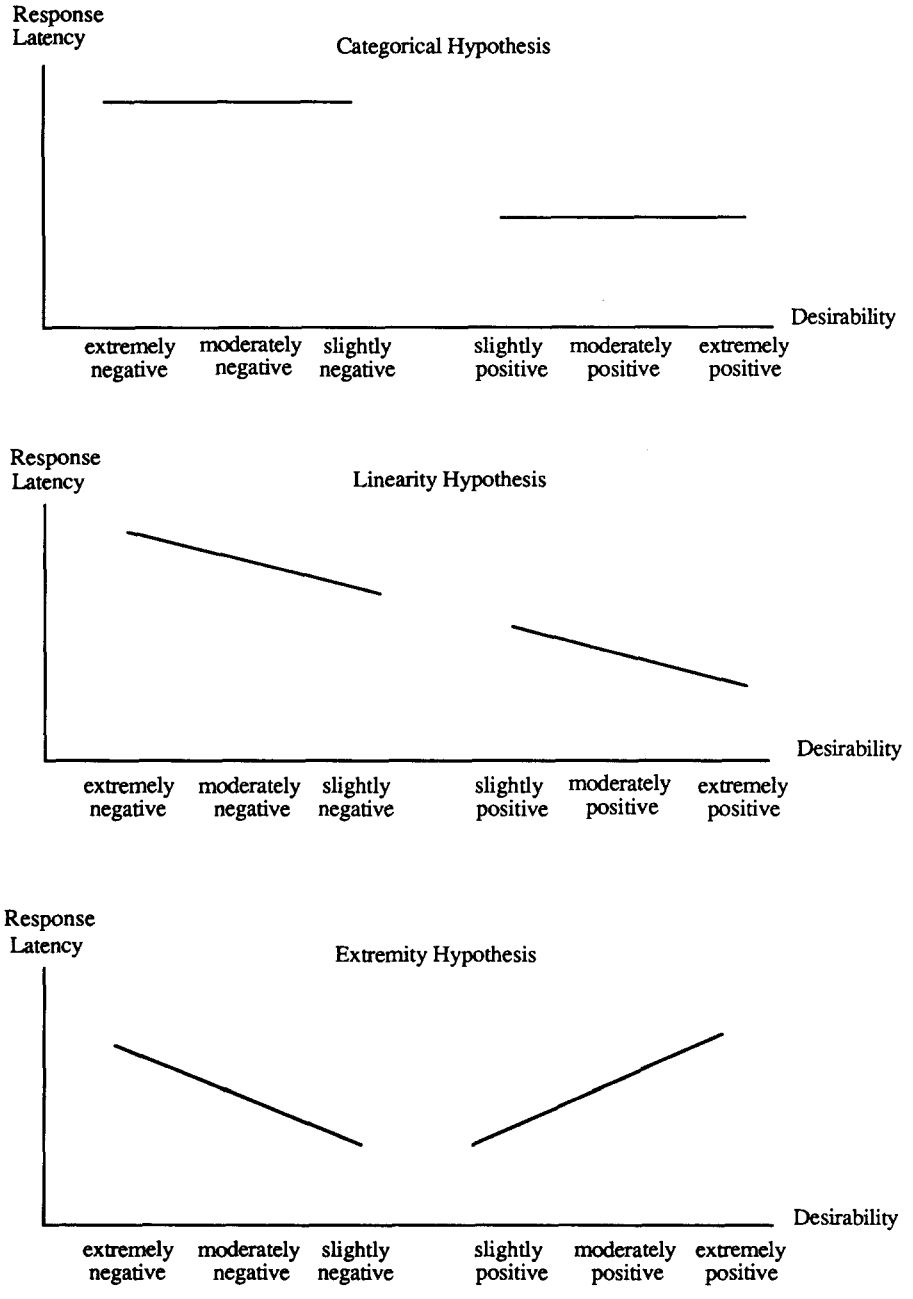

Figure 1. Categorical, linear, and curvilinear relations between color-naming latencies (reaction time in ms) and desirability scale values.

subjects who indicated that they did not learn English before 5 years of age were excluded from all analyses.[1] No color-blind subjects were included in any of the experiments.

Design and stimuli. In the color-naming task, each subject named the colors in which 40 desirable and 40 undesirable personality-trait adjectives were presented; trait valence was thus a within-subjects factor. To achieve a broad and fairly representative coverage of the domain of commonly used English trait adjectives, traits were drawn from each of the Big Five domains of personality description (Norman, 1963; see John, 1990, for a review). We assessed the social desirability of the traits using Hampson, Goldberg, and John's (1987) instructions; 10 judges rated the 80 adjectives on a scale ranging from *extremely undesirable* (1) through *neutral* (5) to *extremely desirable* (9). The mean ratings were highly reliable (Cronbach's coefficient alpha = .98); the mean of the pairwise correlations among the judges was .85, indicating that gradations in trait desirability were highly reliable and that the mean ratings would closely approximate the personal evaluations of most subjects.

[1] Subjects who did not learn English as their first language (e.g., foreign students, immigrants) tend to be slower at color naming; thus, we did not include them in any data analyses.

Indeed, the mean ratings were indistinguishable from those obtained earlier by Hampson et al. (1987); across the 67 adjectives included in both studies, the two sets of mean ratings correlated .96.

Across the mean ratings of the 80 adjectives, the mean was exactly at the scale midpoint of 5.0 ($SD = 2.5$). The mean values spanned almost the entire range of the scale (i.e., from 1.1 to 8.6); *sadistic, mean,* and *hostile* were the most undesirable, and *kind, sincere,* and *talented* were among the most desirable traits. The distribution, however, was bimodal, reflecting the distribution in a comprehensive set of English trait adjectives (Goldberg, 1982); that is, the vast majority of traits are either positively or negatively evaluated. The mean desirability ratings were 7.3 for the 40 desirable traits and 2.7 for the 40 undesirable traits. There was no difference in extremity (i.e., the absolute value of the distance from the scale midpoint of 5.0) between the undesirable traits ($M = 2.30$) and the desirable traits ($M = 2.28$), $t(78) < 1$, and extremity and desirability were uncorrelated ($r = -.08$). Thus, valence and extremity were independent. Word length was counterbalanced between desirable and undesirable trait adjectives; the average number of letters was 7.4 for undesirable adjectives and 7.6 for desirable adjectives.

To permit tests of the linearity and extremity hypotheses in trend analyses, we divided the traits into six categories of desirability values (scale values in parentheses): *extremely undesirable* (range from 1 to 2, $M = 1.5$, $n = 11$ traits), *moderately undesirable* (range from 2 to 3, $M = 2.5$, $n = 16$), *slightly undesirable* (range from 3 to 4, $M = 3.5$, $n = 8$), *slightly desirable* (range from 6 to 7, $M = 6.4$, $n = 12$), *moderately desirable* (range from 7 to 8, $M = 7.4$, $n = 13$) and *extremely desirable* (range from 8 to 9, $M = 8.3$, $n = 12$).[2] These six categories differed significantly from each other in their desirability values (all pairwise $ps < .01$) but did not differ in word length or word frequency.

Apparatus. The experimenter presented instructions and stimuli to subjects on an IBM-PC computer with an EGA color board and monitor running a program in Micro-Experimental Lab. A voice key triggered by microphone input communicated with the software clock through the computer's printer port. Subjects were seated at such a distance that all stimulus words would fall within the foveal area (e.g., Rayner, 1978).

Procedure. The experimenter told the subjects that they would be participating in a color-naming experiment: On each trial a word would appear in the center of the screen, and their task was to name the color in which the word appeared as quickly and as accurately as possible. Subjects completed 15 practice trials, the first 5 of which illustrated the color names. The first 4 experimental trials served as warm-up trials and were not part of the design. The adjective stayed on the screen until the subject triggered the voice key. The experimenter recorded whether the subject named the correct color and whether there was any reason to disregard the response-latency datum (e.g., the voice key was triggered by a cough). After that, 1 s elapsed before the next adjective appeared.

The 80 adjectives were presented in random order; their colors were chosen randomly from the set of blue, green, gold, pink, and red—with the constraint that the same color was never repeated on two consecutive trials, to avoid bias because of accessibility of the color name. After the first 40 trials, subjects were told to take a short break; the first 4 trials after the break did not include experimental stimuli. After the color-naming task, the subjects were probed for suspicion about the purpose of the experiment, were asked at what age they had learned English, and then were debriefed.

Results and Discussion

In this and the other two experiments, subjects made very few color-naming errors ($M = 0.5$), and these errors always occurred with equal frequency for undesirable and desirable traits. Error trials and trials on which a noise other than the color name triggered the voice key were omitted from the analyses. In addition, response times that noticeably deviated from the distribution (under 300 ms or over 1,500 ms) were omitted. In all, less than 1% of the trials were omitted in Experiment 1, and less than 2% in each of the other two experiments.

Effect of valence. To test whether undesirable traits interfered more with the color-naming task than did desirable traits, we tested the effect of trait valence (desirable vs. undesirable) on response latency in a within-subjects analysis of variance (ANOVA), using subjects' response means aggregated across all valid trials as the dependent variable. The effect of valence was significant and in the expected direction, $F(1, 10) = 19.3$, $p = .001$.[3] Subjects took about 29 ms longer to name the color of undesirable traits ($M = 679$ ms) than that of desirable traits ($M = 650$ ms). The mean latency for undesirable traits was greater than the mean latency for desirable traits for 9 (82%) of the 11 subjects.

Although undesirable traits produced significantly more interference than desirable traits for almost all subjects, none of them indicated during debriefing that undesirable traits were more distracting than desirable ones. In fact, subjects reported that they ignored the words and concentrated on recognizing the colors, as they had been instructed. The valence effect occurred although the subjects did not intend to process the trait terms and although they were not aware of their differential attention to desirable and undesirable traits.

Correlations across the traits. In a second set of analyses, we tested whether valence was the only stimulus characteristic related to color-naming latency. Word length had been controlled experimentally. Moreover, word frequency did not affect response latency; the correlation between the mean color-naming latency for each adjective and its frequency in written American English (Francis & Kucera, 1982) was close to 0 ($r = .09$, *ns*). Thus, neither word length nor word frequency could have caused the valence effect.[4]

Additional correlational analyses examined the linearity and extremity hypotheses. If the valence effect simply reflects a linear association, the Pearson product-moment correlation of latency with the continuous desirability values should exceed the point-biserial correlation with valence, and the correlation should be negative within each of the two valence categories. The extremity effect predicts a negative correlation for undesirable traits (i.e., the more extremely undesirable, the longer the latencies) and a positive correlation for desirable traits (i.e., the more extremely desirable, the longer the latencies).

Congruent with the within-subject ANOVA, valence was re-

[2] Because of the bimodal distribution of desirability values in English, the neutral range of desirability, from 4 to 6 on the 1 to 9 rating scale, was represented by only eight traits. These traits also elicited lower agreement among the desirability rates than the more extremely valenced traits, so we omitted them from the present analyses.

[3] All effects significant in the within-subjects analyses reported in this article were also significant when the error term was computed across trials.

[4] Independent of valence, longer words interfered less with color naming than shorter words, $r(78) = -.36$, $p < .01$. The findings for word length and word frequency were the same in all three experiments; these two parameters are therefore not discussed further.

lated to the mean response latencies (point-biserial $r = -.23$, $p < .05$); the Pearson correlation between the desirability ratings and mean latency was also $-.23$ ($p < .05$). That is, the use of continuous desirability values (as opposed to the two valence categories) did not increase the association between latency and negativity. More important, both the linearity and the extremity predictions were contradicted by the two desirability–latency correlations within each valence category: Neither correlation was significantly different from 0, and among the undesirable traits the correlation was positive.

Trend analyses. Linear and curvilinear effects were also tested in a series of trend analyses, with desirability as a within-subjects factor. The mean desirability values of the six desirability categories were used as coefficients in the trend analyses (see Keppel & Zedeck, 1989). The mean color-naming latencies are given for each desirability category in the bottom panel of Figure 2. Neither the linear trend nor the quadratic trend (representing the extremity hypothesis) was significant (both Fs < 1). The valence contrast effect, however, was significant, $F(11, 54) = 31.3$, $p < .001$, and its regression coefficient was significantly different from 0, $t(1) = 3.8$, $p < .001$.

In conclusion, the results of Experiment 1 show that undesirable traits interfered more with the color-naming task than did desirable traits. Both in trend analyses (across subjects) and in correlational analyses (across traits), we found no support for either a linear or an extremity effect of desirability on latencies, suggesting that the effect is categorical. Moreover, this effect cannot be explained by word length or word frequency. Our findings thus confirm previous findings that the evaluation of social stimuli is processed automatically. More important, they support our hypothesis of an asymmetry in the automatic processing of evaluation: The unattended occurrence of an undesirable stimulus interfered more with a primary task requiring attentional resources than did the occurrence of a desirable stimulus.

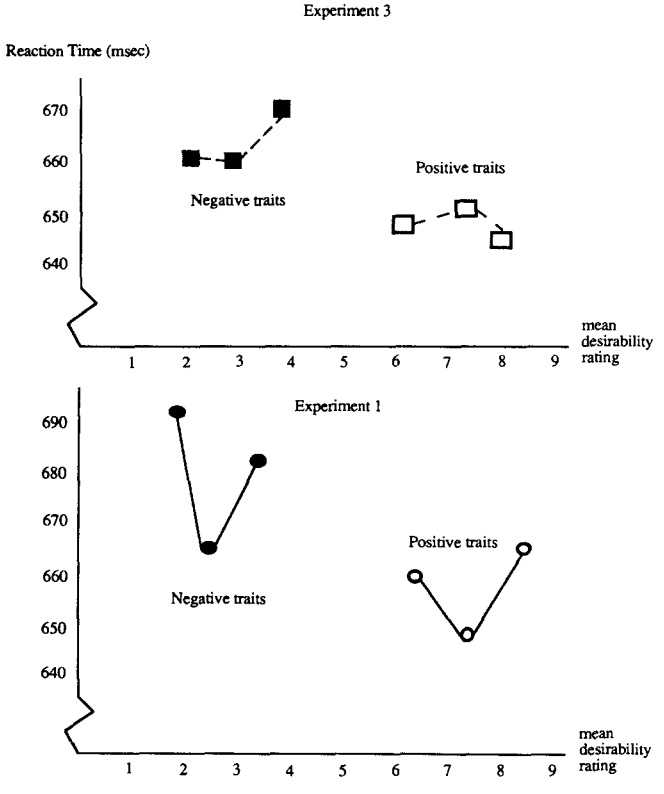

Figure 2. Relation between color-naming latencies (reaction time) and desirability scale values in Experiment 1 (bottom panel) and Experiment 3 (top panel).

Experiment 2: Vigilance and Defense in Incidental Learning

In principle, the longer time subjects needed to respond to the undesirable traits could be due to two quite different mechanisms. Our account postulates perceptual vigilance: Undesirable traits require more time in the color-naming task because negatively valenced stimuli are automatically attended. In general, the material that subjects attend to during presentation will be recalled better (e.g., Fisk & Schneider, 1984). If attention is indeed diverted away from the color-naming task to undesirable traits, as the vigilance account suggests, some incidental learning of these traits should occur, and recall should be greater for undesirable than for desirable traits.

Finding superior incidental recall for undesirable traits would not only strengthen the directed-attention mechanism proposed here, but it would also rule out an alternative, perceptual defense explanation: The color responses to the undesirable traits may have been slower because cognitive effort was required to keep their undesirable content out of consciousness. The notion of defensiveness implies a process motivated by the need to avoid the disturbing affect associated with particular stimuli or memories (see Holmes, 1974, for a review). One type of repression, called *primal repression* or *perceptual defense,* implies that threatening material is kept from entering consciousness (Holmes, 1974, p. 633). A second type, repression proper, suggests that after the material has been consciously recognized, it is relegated to the unconscious. As the stimuli in our color-naming task are not presented subliminally, either type of defense could be involved. Nonetheless, both types lead to the same prediction: The more threatening (undesirable) material should be particularly difficult to retrieve from memory. This line of reasoning is based on the assumption that at least some of the undesirable traits are disturbing or "ego threatening." Therefore, Experiment 2 does not provide a test of whether repression can occur but whether it is responsible for the longer latencies of undesirable traits.

In Experiment 2, then, we tested the vigilance and defensiveness accounts of the negativity effect by comparing the incidental learning of undesirable and desirable traits. Incidental learning was measured by free recall directly following the presentation of desirable and undesirable traits in the color-naming task. However, memory for stimulus aspects unrelated to an attended task tends to be minimal (e.g., Fisk & Schneider, 1984). For example, Bargh and Pratto's (1986) subjects, who named the colors of 50 common noun and trait words, recalled less than 10% of the stimulus words. We therefore modified the design of Experiment 1 in ways that would increase incidental learning. In particular, we presented each adjective twice, used only 40 of

the 80 adjectives from Experiment 1, and to compensate for the resulting loss in power, we doubled the number of subjects. Thus, Experiment 2 provides a replication of the color-interference effect with a less extensive set of traits, and the repeated presentation of the traits allows us to examine whether this effect is influenced by habituation and practice.

Method

Subjects. Subjects were 32 undergraduates from the University of California, Berkeley, who volunteered to participate and received partial course credit. Data from 3 subjects who had learned English after age 5 and from 4 additional subjects who indicated during debriefing that they had expected the incidental-recall task were omitted from the analyses, leaving a total of 25 subjects.

Trait stimuli. To ensure that the 40 traits included a similar range and diversity of content as the initial set, the 80 traits were grouped into 40 pairs of quasi synonyms (e.g., *sadistic* and *mean*), and only one of the synonyms was included in the abbreviated set. As in Experiment 1, word length was controlled, and the 20 desirable and 20 undesirable traits differed significantly in desirability but not in extremity. The 40 traits were presented in two different random orders: To control for primacy and recency effects on recall, one order began with two desirable traits and ended with two undesirable ones, whereas the other had the opposite pattern. Subjects received both orders, in either Block 1 or Block 2, and the assignment of orders to blocks was counterbalanced across subjects. The five colors were counterbalanced across desirable and undesirable traits, and the same color was never presented in consecutive trials. Within these constraints, colors were assigned randomly to the traits.

Procedure. The color-naming part of the experiment followed the same procedure as in Experiment 1, except that the color name *yellow* was used instead of *gold.* All 80 trials were presented consecutively without a break. Immediately after the last color-naming trial, instructions appeared on the computer screen asking the subjects to write down all of the words they could remember on a blank sheet of paper. When subjects could not recall any more words, they were interviewed about the task and then were debriefed.

Results

Color-naming latencies. Because the same 40 traits had been presented twice in two different orders in the two blocks, we were able to test the joint effects of valence and block on the response latencies in a 2 × 2 within-subjects ANOVA, using as the dependent variable each subject's mean response time for undesirable and for desirable traits in each of the two blocks. As in Experiment 1, there was a significant main effect for valence, $F(1, 24) = 9.1$, $p = .006$. Again, the mean response latencies were longer for undesirable traits ($M = 612$ ms) than for desirable ones ($M = 601$), and this effect held for 19 (76%) of the 25 subjects. As shown in Figure 3, subjects were slightly faster in the second block, but neither the main effect of block, $F(1, 24) = 1.6$, $p = .22$, nor the interaction ($F < 1$) was significant. Neither the two orders in which the traits had been presented nor their assignment to Block 1 or 2 had any effect on the color-naming latencies. Regardless of block and order, then, the present findings replicated those of Experiment 1.

Correlations across the traits. The power of correlational (and trend) analyses is limited by the small number of traits presented. Nonetheless, the pattern of correlations closely mirrored the categorical pattern found in Experiment 1. We again found the negative point-biserial correlation between response latency and valence ($r = -.24$, $p < .05$, one-tailed), and that correlation was larger than the Pearson correlation between latency and the graded desirability values ($r = -.18$, *ns*). Within the two valence categories, the correlations between latency and desirability were almost exactly 0.

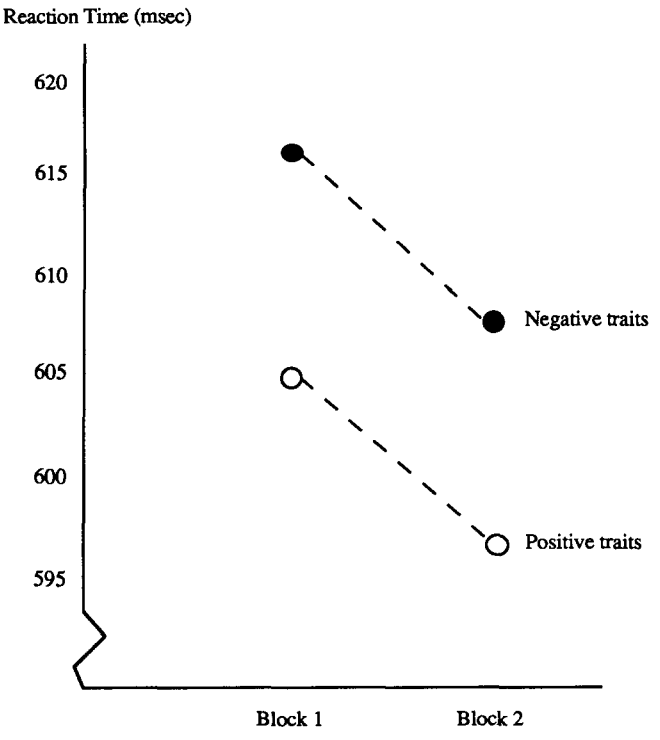

Figure 3. Color-naming latencies (reaction times) by valence and block in Experiment 2.

Free recall. Subjects recalled, on the average, only 3.9 traits, with a range from 0 to 7; subjects expressed surprise at being asked to recall the words, and most apologized for being able to recall so few. They did, however, recall twice as many undesirable ($M = 2.6$, range = 0 to 5) as desirable ($M = 1.3$, range = 0 to 3) traits. The difference in the number of undesirable and desirable words recalled by each subject was significantly different from 0, $t(24) = 3.4$, $p < .01$.

Individual-differences analyses showed a substantial floor effect associated with a low level of recall; subjects recalling more traits overall were likely to show a more pronounced difference between undesirable and desirable traits, $r(23) = .50$, $p < .01$. Nonetheless, the superior recall of undesirable traits held for almost two thirds of the subjects; 16 recalled more undesirable than desirable traits, 8 recalled equal numbers of undesirable and desirable traits (including the 1 subject who recalled none), and 1 recalled more desirable than undesirable traits.[5]

[5] Of the 7 excluded subjects who had learned English after the age of 5 or expected the incidental recall task, 4 showed the valence effect and 3 did not. When their data are included, the difference is still significant, $t(31) = 2.5$, $p < .02$.

Analyses across the 40 traits showed that all 20 of the undesirable traits had been recalled by at least 1 subject, whereas only 15 of the desirable traits had been recalled by at least 1 subject. That is, the valence effect on recall was not due to the superior recall of a small set of undesirable traits. Across the 40 traits, the correlation between valence and the number of subjects who recalled the trait was $-.31$ ($p < .05$). The correlation between social desirability and the number of subjects who recalled the trait was $-.39$ ($p < .05$), which did not reliably differ from the point-biserial correlation ($p = .29$).

Response latency, recency of exposure, and recall. We have argued that longer color-naming latencies indicate greater attention to a stimulus and that the stimuli to which subjects pay more attention should be better recalled. If this is true, the mean color-naming latency for a word should be positively related to its frequency of recall. Indeed, across the 40 traits, the correlation between the mean response latency and the number of subjects recalling the trait was positive. This positive association was more pronounced for response latencies measured immediately before recall, that is, in the second block ($r = .31$, $p < .05$), than for latencies measured in the first block ($r = .16$, *ns*), suggesting a recency effect; the most recent presentation of a word is the more potent determinant of recall. However, this recency effect was not particularly strong; when frequency of recall was computed separately for subjects receiving Order A and for those receiving Order B, in the second (more recent) block, the correlation between the resulting two measures of recall was .45 ($p < .01$) across the 40 words, indicating that the valence of the trait stimulus determined its recall, rather than the order and recency of its presentation.

Discussion

The results of Experiments 1 and 2 showed that undesirable traits are more likely than desirable traits to attract attention even when attention is deliberately focused elsewhere. The latency difference established in Experiment 1 was replicated in Experiment 2 in both the first and second stimulus presentations. Moreover, the correlations between desirability and latency followed the same categorical pattern as in the first study.

The free-recall data suggest two major conclusions. First, unintentional processing of the meaning of the trait stimuli produced very little memory. The "best" subject recalled only 7 words, and the average subject recalled less than 10% of the 40 stimuli, each of which had been presented twice. The low rate of recall is consistent with the short response latencies in the color-naming task, both suggesting that subjects focused their attention on the colors of the terms, not on their meanings. In fact, some subjects did not even realize that the stimuli had been adjectives.

Second, we found that the longer color-naming latencies for undesirable traits were associated with greater accessibility in memory. Across subjects, undesirable traits were recalled twice as often as desirable traits; across traits, there were twice as many subjects showing superior recall for undesirable traits as there were subjects showing no differential recall. These findings rule out the defensiveness hypothesis and, more important, provide overwhelming support for the vigilance interpretation. Undesirable traits require longer response times in the color-naming paradigm not because cognitive work is required to shut out the perception of such negative stimuli or to relegate them to the unconscious once recognized. Rather, our findings are most consistent with the hypothesis that undesirable traits automatically attract more attention and are therefore better remembered than desirable traits. The positive correlation between mean response latency and frequency of recall is consistent with the assumption that differential attention influences recall.

Experiment 3: Automatic Processing of Base Rate and Valence

The findings of Experiment 2 strengthen the automatic vigilance hypothesis. However, in studies of impression formation and person perception (e.g., Anderson, 1974; Fiske, 1980; see Skowronski & Carlston, 1989; Taylor & Fiske, 1978, for reviews), the stronger weighing of negative than of positive information has been explained in informational terms by the higher informativeness (or diagnosticity) of negative information. Negative information tends to be perceived as more diagnostic than positive information because people's expectations about events and outcomes in the world are generally positive. For example, people expect others to behave in socially desirable or at least socially appropriate ways (Kanouse & Hanson, 1972). According to the widely demonstrated positivity effect in person perception (e.g., Sears, 1983), people assume that most individuals have desirable characteristics. The Pollyanna principle (Matlin & Stang, 1978) suggests that people expect positive outcomes even when faced with information to the contrary. In other words, desirable events tend to be viewed as common, frequently occurring, and typical, whereas undesirable events tend to be seen as uncommon, infrequent, and atypical. The informational value of undesirable traits should be higher than that of desirable traits, as uncommon and atypical events are seen as more informative (see Fiske, 1980) and diagnostic (Skowronski & Carlston, 1989; see also Lay, Burron, & Jackson, 1973).

Rating studies have shown a positive and substantial relation between the desirability of personality traits and their perceived base rate (or frequency in the population): The more desirable the trait, the more frequent is it perceived to be (e.g., Fulero, 1979; Funder & Dobroth, 1987; Rothbart & Park, 1986). If this relation holds in the present studies, the lower base rate of undesirable information might account for the negativity effect. To test this possibility, we obtained estimates of desirability and perceived base rate from additional groups of subjects and conducted a third color-naming experiment to test whether our earlier findings are best interpreted in terms of valence, base rate, or both.

An experimental comparison of the valence and the infrequency hypotheses requires trait stimuli for which one hypothesis predicts interference but the other does not: desirable traits considered infrequent and undesirable traits considered frequent. The 80 traits used in Experiment 1 had been selected without prior consideration of their perceived base rates. To this initial set, we added another 51 traits to represent the two conditions needed to unconfound valence from base rate. We obtained both desirability and base-rate ratings and con-

structed a set of trait stimuli to manipulate valence and base rate independently. All 131 traits were then used as stimuli in the color-naming task, thus permitting us to replicate earlier analyses with the original set of 80 traits, to test the joint effects of valence and base rate in an unconfounded set of 88 traits, and to replicate the linear and quadratic trend analyses in a large set of stimuli, consisting of all 131 traits.

Method

Base-rate and desirability ratings for 131 traits. The 51 additional traits were selected from previous rating studies by Fulero (1979), Funder and Dobroth (1987), Hampson et al. (1987), Norman (1967), and Rothbart and Park (1986) or were newly generated on the basis of these studies.[6] Roughly one half of the additional traits were expected to be undesirable but common (low valence, high base rate) and the other half desirable but uncommon (high valence, low base rate).

We drew two samples from the same population as the experimental subjects. One sample ($n = 12$) rated all 131 traits on the nine-step social desirability scale used in Experiment 1, and the other ($n = 16$) estimated the base rate of each of the traits. The base-rate judges were asked to round their estimates to the nearest 5%. Part of their instructions read as follows:

> What percentage of people can be characterized by a particular personality trait? In this study, we are trying to discover the "base rate"—that is, the relative frequency—of each of a number of personality characteristics in the general population. . . . Consider only those persons who are of the same sex as you are and of your approximate age. Of such persons, please indicate the *percentage* that are characterized by each of the following traits.

Both types of ratings proved highly reliable across judges (both Cronbach coefficient alphas = .98). The mean desirability values ranged from 1.3 (*bigoted*) to 8.5 (*honest*), with a mean of 5.2 ($SD = 2.3$) and correlated .96 with our earlier ratings across the common set of 80 traits. The mean base-rate values ranged from 7% (*saintly*) to 78% (*curious*), with a mean of 42% ($SD = 15\%$); in contrast to the desirability ratings, they had a unimodal distribution.

Across the 80 traits studied previously, the base rates correlated .43 with the original desirability ratings, .34 with the new desirability ratings, and .30 with valence (all $ps < .01$). In contrast, across all 131 traits, the correlations were not significant, with correlations of −.12 for desirability and −.14 for valence and with correlations of zero within the subsets of desirable and undesirable traits. That is, valence and base rate were confounded in the initial set of traits but not in the new set.

Manipulation of valence and base rate. From the 131 traits, 88 were selected to form a 2 × 2 Valence × Base Rate design with 22 traits per condition. These 88 traits in the 4 Valence × Base Rate cells, as well as the mean ratings of the traits in the 4 cells, are listed in the Appendix. The traits classified as common ranged from 44% to 78%; those classified as uncommon ranged from 7% to 42%. For the mean desirability ratings, undesirable traits ranged from 1.3 to 4.0, and desirable traits ranged from 6.3 to 8.5. An ANOVA on the mean base-rate ratings yielded a significant effect for base-rate condition, $F(1, 84) = 247$, but not for valence ($F < 1$) or their interaction ($F = 1.2$). Conversely, an ANOVA on the mean desirability ratings yielded a significant effect for valence, $F(1, 84) = 1{,}333$, but not for base-rate condition or their interaction (both $Fs < 1$). Thus, base rate and valence were completely unconfounded in this design.

Subjects and procedure in the color-naming task. Subjects were 17 Berkeley undergraduates, who volunteered to participate in exchange for partial course credit. The 131 traits were presented in one random order; otherwise, the procedures were the same as in Experiment 1, except that the color name *yellow* was used instead of *gold*.

Results and Discussion

Joint effects of base rate and valence. The mean of the 22 traits in each condition served as the dependent variable in a 2 (valence) × 2 (base rate) within-subjects ANOVA. As shown in Figure 4, the main effect of valence was significant, $F(1, 16) = 52.5$, $p < .001$; once again, subjects took longer to name the color of undesirable traits ($M = 676$ ms) than of desirable traits ($M = 647$ ms), and all 17 subjects showed this effect. In contrast, both the main effect of base rate and its interaction with valence were not significant, $Fs(1, 16) = 3.2$ and 1.6, respectively. More important, the direction of the base-rate effect was opposite to expectations; as shown in Figure 4, common traits elicited slightly longer latencies ($M = 667$ ms) than uncommon traits ($M = 656$ ms). This insignificant difference was due to the common undesirable traits, which elicited the longest response latencies.

In a second ANOVA, the 2 (valence) × 2 (base rate) design was analyzed across the 88 traits, using the mean response latency across subjects as the dependent variable and each trait as an observation. This ANOVA also showed a significant valence effect, $F(1, 84) = 11.2$, $p = .001$, and neither a main effect of base rate ($p = .21$) nor a Base Rate × Valence interaction ($p = .27$).

Base-rate effects in the traits used in Experiments 1 and 2. To examine the generalizability of these findings, we used median splits to divide the 80 traits studied in Experiment 1 into a 2 (valence) × 2 (base-rate conditions) design and subjected the mean response latencies to each trait to an unbalanced ANOVA across traits. As in Experiment 3, the valence effect was significant, $F(1, 76) = 4.2$, $p < .05$, and there was neither a main effect of base rate nor an interaction (both $Fs < 1$), despite the fact that valence and base rate were positively correlated.

The same picture emerged in a series of correlational analyses. In particular, the correlations between the mean response latencies and base rate were not significant in all three studies: for the 80 traits in Experiment 1 ($r = -.14$), the 40 traits in Experiment 2 ($r = .02$), and the 131 traits in Experiment 3 ($r = -.03$). These findings provide no support for a base-rate interpretation of the negativity effect; in fact, in the 88 traits in Experiment 3 for which base rate and valence were independent, the size of the valence effect was just as strong (29 ms) as in Experiment 1.

Categorical versus linear and quadratic effects on latency. In another test of the linear and extremity hypotheses, we performed trend analyses analogous to those in Experiment 1, using the combined set of traits to create desirability categories. Again, we used six categories, which differed pairwise in their mean desirability ratings (all $ps < .01$). As in Experiment 1, the linear and quadratic trends were all insignificant (all $Fs < 1$). Only the valence contrast was significant; the overall contrast effect was $F(17, 84) = 71.3$, $p = .001$, and for the regression weight of the valence variable, $t(1) = 3.0$, $p < .005$. Across the 131 traits in this analysis, 15 of the 17 subjects (88%) showed the negativity effect. When the data from Experiments 1 (bottom panel of Figure 2) and Experiment 3 (top panel of Figure 3) are

[6] We are indebted to David Funder and Myron Rothbart for providing us with base-rate ratings and to Eileen Donahue and Delroy Paulhus for their imaginativeness in helping us generate additional traits.

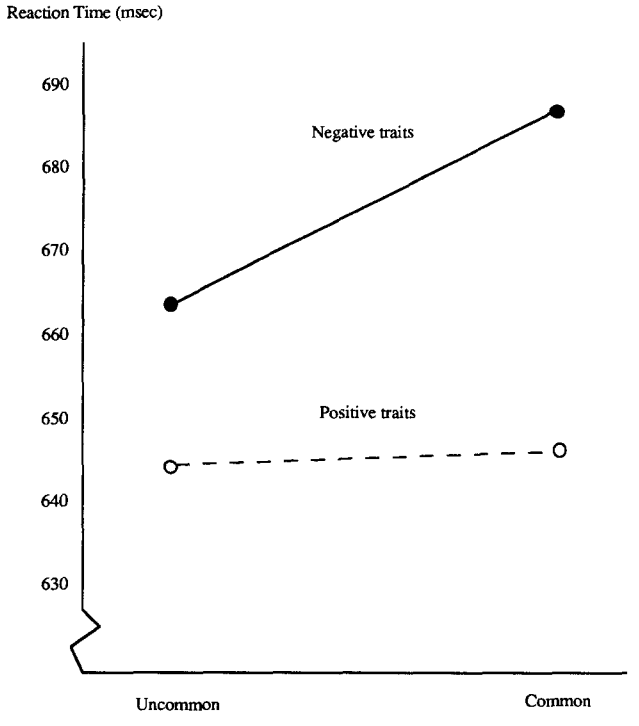

Figure 4. Color-naming latencies (reaction times) by valence and base-rate condition in Experiment 3.

considered together, there is little evidence for any consistent trends in the data except the categorical effect of valence.[7]

These conclusions are borne out by the correlational analyses, regardless of the experiment or the subset of words analyzed. In the set of 88 traits for which we selected only clearly desirable and clearly undesirable traits, the correlation of response latency and valence was .34 ($p < .01$), as contrasted with .11 among the desirable traits and −.14 among the undesirable traits. For all 131 traits, the correlation with valence was still −.17 ($p < .05$), as contrasted with −.03 and .02 for the desirable and undesirable subsets, respectively. These correlations are very similar to those in the other two experiments and show a significant effect of valence but no evidence for any linear or curvilinear effects of the graded desirability values.

General Discussion

One purpose of automatic evaluation, we have argued, is to direct attention to negatively evaluated stimuli, and this shift in attention occurs without the perceiver's intent. In each of three experiments, we found that undesirable traits attracted more attention (as reflected in response latencies) than desirable traits; combined across studies, 85% of all subjects showed the valence effect. The difference between undesirable and desirable traits was 29 ms in Experiments 1 and 3; in Experiment 2, which used fewer stimuli, the size of the effect was somewhat smaller. Neither word length nor word frequency mediated the effect in any of the experiments.

Moreover, we ruled out two of the most plausible alternative explanations. The incidental-learning findings in Experiment 2 are inconsistent with a perceptual-defense interpretation of the valence effect and instead strengthen the vigilance hypothesis: The interference in the color-naming task was associated with better recall, supporting the hypothesis that more attentional resources were devoted to undesirable stimuli.

In Experiment 3, we manipulated both the valence and the perceived base rate of the traits. Whereas the valence effect was exactly replicated, the effect of base rate was not significant and not consistent with the hypothesis that infrequent traits attract more attention. Our reanalysis of the data from Experiment 1 and the correlational analyses relating base rate to mean latencies in all three experiments led to the same conclusion: The valence effect cannot be explained by base-rate differences among the traits, and base rate does not seem to play an important role in the color-naming paradigm. Taken together, the results of the three experiments provide converging evidence for the hypothesis that people's attention is drawn to negative information without their intention and that the cause of this effect lies in the valence of the traits, not their informational value or diagnosticity.

The Categorical Nature of the Valence Effect

In all three studies, we examined the data for possible linear and quadratic trends, using both correlational analyses (across stimuli) and within-subjects trend analyses. In all of these analyses, we found no support for any effect but the categorical one. The absence of linear and quadratic effects cannot be explained by the reliability of the desirability scale values because these effects failed to emerge even in the within-subjects trend analyses involving highly reliable contrasts, such as those between Desirability Categories 1 and 3 among the undesirable traits and between Categories 4 and 6 among the desirable traits. This finding is particularly surprising because the power of linear models, even improper ones, has been well demonstrated (Dawes, 1979). Linear models approximate categorical effects, which are therefore more difficult to establish than linear ones; it is extremely rare for a Pearson correlation, which is based on a continuous variable (e.g., desirability) to be eclipsed by the point-biserial correlation, which is based on a dichotomized measure of the same variable (e.g., valence).

Nonetheless, the finding that undesirable traits, regardless of their extremity, are more likely to attract attention than desirable traits requires further comment. In particular, the evolutionary argument (that it is adaptive to attend to stimuli associated with negative consequences) might be misunderstood to mean that only specific subclasses of undesirable traits should show attention-grabbing effects, namely, only those traits of others that directly endanger our personal well-being, such as *sadistic* or *violent*. However, this interpretation confuses distal (evolutionary) with proximal (psychological) mechanisms and automatic processing with controlled processing.

In particular, we postulated that an adaptive behavior (i.e.,

[7] Because different sets of traits were used in the two experiments, the mean desirability values of the six desirability categories differed slightly across experiments.

monitoring the environment for potential danger) is accomplished through a psychological mechanism, which we have called *automatic vigilance.* This mechanism is not contingent on the physical presence of another person posing a threat; after all, our effects were obtained with colored words presented on a computer screen. Moreover, given its speed, we did not expect the attention-grabbing effect to be very differentiated. Automatic vigilance functions as a signal, rather than by providing a detailed analysis of the stimulus. Indeed, as we argue below, linear and extremity effects are more likely to result from controlled processing in deliberate evaluative judgments, such as desirability ratings and impression formation.

Automatic Evaluation and Automatic Vigilance

Given the consistency of our results across studies, one may wonder why this effect was not discovered in previous studies on automatic evaluation. Fazio et al. (1986), for example, did not find significant differences between positively and negatively valenced attitude objects. However, being interested in demonstrating the automatic evaluation effect, not a negative–positive asymmetry, they used only a few stimuli, which were selected idiographically for each subject. Our findings suggest that relatively large and systematic selections of stimuli may be necessary to demonstrate the automatic vigilance effect. In Experiment 2, which used fewer stimuli, we found that the effect was smaller than in the two other experiments.

A second difference between the present and past studies is the nature of the task. An interference task, such as the color-naming task, is the ideal paradigm to study automatic vigilance. This paradigm mimics in the laboratory a real-life setting in which the subject is concerned with other activities (i.e., the attended task) while automatically monitoring the perceptual field for undesirable events or stimuli. In this kind of situation, automatic vigilance is important because it can redirect attention to information about potentially undesirable events. In settings in which redirection of attention is not possible or not necessary, however, automatic vigilance is irrelevant. For example, when subjects are already attending to the evaluation of the stimulus, or when the response mode makes the evaluation of the stimuli salient, automatic vigilance effects may not be observed. The automatic vigilance mechanism does not imply that undesirable stimuli are necessarily recognized faster or more accurately than desirable ones; rather, when attentional resources are directed elsewhere, undesirable stimuli are more likely to attract such resources.

What happens once an undesirable stimulus has attracted attention? How does that attention influence subsequent processing? The answer to these questions depends on the particular task the perceiver is trying to accomplish. In the color-naming task, vigilance was unwarranted and interfered with subjects' goals. Note that although undesirable traits elicited longer response latencies, they did not cause more color-naming errors, nor did the subjects necessarily become conscious of the attentional shift to the evaluative aspects of the stimulus. Thus, just as subjects can inhibit the word meaning in the Stroop (1935) interference paradigm, they were here able to control the automatically grabbed attentional resources, so that the intrusive effects of automatic vigilance were not noticed.

Unintended Effects of Automatic Processing on Intentional Processing

However, the color-naming task is unusual in that automatic vigilance is rendered inappropriate, even dysfunctional. In more typical contexts, the attention shifted to an undesirable stimulus permits more deliberate and controlled processing of that stimulus. In general, attentional focus has been shown to influence social judgment (see Taylor & Fiske, 1978), and automatic vigilance can thus lead to negative bias in judgment.

A second link between automatic vigilance and social judgment is suggested by our second experiment. Even in the color-naming paradigm, the greater attentional resources allocated to undesirable traits increased subsequent memory for these traits. Automatic vigilance made these undesirable stimuli relatively more accessible in memory. Judgment can be heavily influenced by the information accessible in memory; therefore negatively biased memory may contribute to negatively biased judgments. This link, relating automatic vigilance to biased memory and, in turn, biased memory to biased judgment, may provide an important theoretical connection between the present research and earlier research on the differential weighing of desirable and undesirable information about others.

In particular, the link between vigilance, memory, and judgment would predict greater weighing of undesirable stimuli in impression formation tasks. For example, if the traits whose colors the subjects named were a description of an individual or a social group, our finding that subjects remembered twice as many undesirable than desirable traits (although an equal number was presented) would predict that subjects' impressions of that individual or group would be negatively biased. More generally, automatic vigilance is a mechanism that could explain why unfavorable information about individuals and stereotyped groups is often noticed and remembered better than favorable information, even when the social perceiver is not intentionally processing this information.

In summary, our findings and the close associations between attention, memory, and judgment suggest that automatic vigilance alone could lead to the differential weighing of undesirable information that is typically observed in impression formation studies. However, additional processes must be involved because the differential weighing effects seem considerably more complex than the automatic effects demonstrated in the present research. In particular, in her study of looking time and weighing of desirable and undesirable behaviors, Fiske (1980) found that very undesirable behaviors differed from somewhat undesirable behaviors, whereas we found no such linear effect. Moreover, very desirable behaviors elicited longer looking times and greater weights than did somewhat desirable traits, an extremity effect we did not obtain. Finally, Fiske (1980) explained her findings in terms of informativeness, assuming that regardless of their valence, extreme behaviors are more informative than less extreme behaviors, whereas we ruled out infrequency as an explanation of the automatic vigilance effect.[8]

[8] Fiske (1980) manipulated the desirability of the behaviors presented on the slides but did not measure their perceived frequency (or base rate).

The difference, we suppose, lies in the deliberate and controlled processing that occurred in Fiske's (1980) impression formation task: Subjects were instructed to examine the information presented as long as they wished and to form an impression of the target person. Such intentional processing was prevented in the color-naming task. The linear and extremity effects, observed in impression formation but not in the color-naming task, thus seem to depend on intentional processing. Once attention has been directed at negative information, the subsequent use and weighing of this information during deliberate processing may depend on a number of factors, including informativeness or diagnosticity (see Skowronski & Carlston, 1989).

In conclusion, we view automatic vigilance as a "default" response: It monitors potentially undesirable information when specific impression formation goals are not active, and it serves as an input to deliberate processing when such goals are (or have become) active. In principle, the effects of automatic vigilance can be overridden by other goals (as in the color-naming task), although the presence of negativity bias in our incidental-learning study and in the numerous impression formation studies suggests that these conditions are unlikely to completely eliminate its effect. However, when perceivers can determine what information is made available to them (as in interviews), the goal to be accurate can lead them to be less biased in seeking negative information and to form less negatively biased impressions, even when they have negative expectancies about the target (Neuberg, 1989). Thus, bias toward undesirable information and the influence of such information on judgment seem most pronounced when people do not realize that such influences are occurring or when they are not motivated to prevent them from occurring (see Bargh, 1989). These two conditions probably hold for most situations in which intergroup contact occurs and beliefs about out-groups are formed and confirmed. In these situations, automatic vigilance might foster the formation and maintenance of unfavorable impressions and stereotypes. Thus, people's greater attention to negative information may protect them from immediate harm but it may also contribute to prejudice and conflict in social interaction.

References

Anderson, N. H. (1974). Cognitive algebra: Integration theory applied to social attribution. In L. Berkowitz (Ed.), *Advances in experimental social psychology,* (Vol. 7, pp. 2–102). San Diego, CA: Academic Press.

Bargh, J. A. (1989). Conditional automaticity: Varieties of automatic influence in social perception and cognition. In J. S. Uleman & J. A. Bargh (Eds.), *Unintended thought: Causes and consequences for judgment, emotion, and behavior* (pp. 3–51). New York: Guilford Press.

Bargh, J. A., Litt, J. E., Pratto, F., & Spielman, L. A. (1988). On the preconscious evaluation of social stimuli. In K. McConkey & A. Bennet (Eds.), *Proceedings of the XXIV International Congress of Psychology* (Vol. 3, pp. 1–57). Amsterdam: Elsevier/North-Holland.

Bargh, J. A., & Pratto, F. (1986). Individual construct accessibility and perceptual selection. *Journal of Experimental Social Psychology, 22,* 293–311.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34,* 571–582.

Fazio, R. H. (1986). How do attitudes guide behavior? In R. M. Sorrentino & E. T. Higgins (Eds.), *The handbook of motivation and cognition: Foundations of social behavior* (pp. 204–243). New York: Guilford Press.

Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology, 50,* 229–238.

Fisk, A. D., & Schneider, W. (1984). Memory as a function of attention, level of processing, and automatization. *Journal of Experimental Psychology: Learning, Memory and Cognition, 10,* 181–197.

Fiske, S. T. (1980). Attention and weight in person perception: The impact of negative and extreme behavior. *Journal of Personality and Social Psychology, 38,* 889–906.

Fiske, S. T. (1982). Schema-triggered affect: Applications to person perception. In M. S. Clark & S. T. Fiske (Eds.), *Affect and cognition* (pp. 55–78). Hillsdale, NJ: Erlbaum.

Francis, W. N., & Kucera, H. (1982). *Frequency analysis of English usage.* Boston: Houghton Mifflin.

Frodi, L. M., Lamb, M. E., Leavitt, L. A., & Donovan, W. L. (1978). Fathers' and mothers' responses to infants' smiles and cries. *Infant Behavior and Development, 1,* 187–198.

Fulero, S. (1979). *Recall of confirming events as a function of favorability and frequency.* Unpublished doctoral dissertation, Department of Psychology, University of Oregon, Eugene.

Funder, D. C., & Dobroth, K. M. (1987). Differences between traits: Properties associated with interjudge agreement. *Journal of Personality and Social Psychology, 52,* 409–418.

Goldberg, L. R. (1982). From Ace to Zombie: Some explorations in the language of personality. In C. D. Spielberger & J. N. Butcher (Eds.), *Advances in personality assessment* (Vol. 1, pp. 203–234). Hillsdale, NJ: Erlbaum.

Hamilton, D. L., & Zanna, M. (1972). Differential weighing of favorable and unfavorable attributes in impressions of personality. *Journal of Experimental Research in Personality, 6,* 204–212.

Hampson, S. E., Goldberg, L. R., & John, O. P. (1987). Category-breadth and social desirability values for 573 personality terms. *European Journal of Personality, 1,* 241–258.

Holmes, D. S. (1974). Investigations of repression: Differential recall of material experimentally or naturally associated with ego threat. *Psychological Bulletin, 81,* 632–653.

John, O. P. (1990). The "Big Five" factor taxonomy: Dimensions of personality in the natural language and in questionnaires. In L. A. Pervin (Ed.), *Handbook of personality: Theory and research* (pp. 66–100). New York: Guilford Press.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47,* 263–291.

Kahneman, D., & Tversky, A. (1984). Choices, values, and frames. *American Psychologist, 39,* 341–350.

Kanouse, D. E., & Hanson, L. R. (1972). Negativity in evaluations. In E. E. Jones, D. Kanouse, H. H. Kelley, R. E. Nisbett, S. Valins, & B. Weiner (Eds.), *Attribution: Perceiving the causes of behavior* (pp. 47–62). Morristown, NJ: General Learning Press.

Keppel, G., & Zedeck, S. (1989). *Data analysis for research designs.* New York: Freeman.

Kunst-Wilson, W. R., & Zajonc, R. B. (1980). Affective discrimination of stimuli that cannot be recognized. *Science, 207,* 557–558.

Lay, C. H., Burron, B. F., & Jackson, D. N. (1973). Base rate and informational value in impression formation. *Journal of Personality and Social Psychology, 28,* 390–395.

Lazarus, R. S. (1966). *Psychological stress and the coping process.* New York: McGraw-Hill.

Lazarus, R. S. (1982). Thoughts on the relations between emotion and cognition. *American Psychologist, 37,* 1019–1024.

Matlin, M., & Stang, D. (1978). *The Pollyanna principle.* Cambridge, MA: Schenkman.

Neuberg, S. L. (1989). The goal of forming accurate impressions during social interactions: Attenuating the impact of negative expectancies. *Journal of Personality and Social Psychology, 56,* 374–386.

Norman, W. T. (1963). Toward an adequate taxonomy of personality

attributes: Replicated factor structure in nomination of personality ratings. *Journal of Abnormal and Social Psychology, 66,* 574–583.

Norman, W. T. (1967, April). *2,800 personality trait descriptors: Normative operating characteristics for a university population* (Tech. Rep.). Ann Arbor, MI: University of Michigan, Department of Psychology.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning.* Urbana: University of Illinois Press.

Rayner, K. (1978). Foveal and parafoveal cues in reading. In J. Requin (Ed.), *Attention and performance VIII* (pp. 149–161). Hillsdale, NJ: Erlbaum.

Rothbart, M., & Park, B. (1986). On the confirmability and disconfirmability of trait concepts. *Journal of Personality and Social Psychology, 50,* 131–142.

Seamon, J. G., Brody, N., & Kauff, D. M. (1983). Affective discrimination of stimuli that are not recognized: Effects of shadowing, masking, and cerebral laterality. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9,* 544–555.

Seamon, J. G., Marsh, R. L., & Brody, N. (1984). Critical importance of exposure duration for affective discrimination of stimuli that are not recognized. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 465–469.

Sears, D. O. (1983). The person–positivity bias. *Journal of Personality and Social Psychology, 44,* 233–250.

Shiffrin, R. M. (1988). Attention. In R. C. Atkinson, R. J. Herrnstein, G. Lindzey, & R. D. Luce (Eds.), *Stevens' handbook of experimental psychology: Vol. 2. Learning and cognition* (pp. 739–811). New York: Wiley.

Skowronski, J. J., & Carlston, D. E. (1989). Negativity and extremity biases in impression formation: A review of explanations. *Psychological Bulletin, 105,* 131–142.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18,* 643–662.

Taylor, S. E., & Fiske, S. T. (1978). Salience, attention, and attribution: Top of the head phenomena. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 11, pp. 249–288). San Diego, CA: Academic Press.

Appendix

The 88 Trait Terms Used in the Base-Rate Study (Experiment 3)

Undesirable		Desirable		Undesirable		Desirable	
Uncommon	Common	Uncommon	Common	Uncommon	Common	Uncommon	Common
rude	bigoted	exact	tolerant	sassy	messy	charming	kind
wicked	selfish	polished	curious	forgetful	gossipy	ingenious	talented
sadistic	irritable	refined	extroverted	glum	stubborn	scholarly	smart
mean	immature	humble	vigorous	curt	contradictory	musical	happy
hostile	tactless	worldly	stable	finicky	fickle	artistic	caring
intolerant	jealous	concise	inquisitive	sad	gullible	inventive	creative
annoying	cranky	saintly	active	passive	insecure	wise	loving
bossy	shallow	dignified	organized	naive	sarcastic	brilliant	honest
stingy	nosy	gracious	confident				
sluggish	biased	cultured	polite		Mean desirability		
lazy	moody	original	reliable				
pesty	boastful	elegant	perceptive	2.6	2.7	7.5	7.5
stupid	wasteful	heroic	helpful		Mean base rate		
domineering	impatient	witty	sincere				
				30%	56%	27%	57%

Note. Within each of the four conditions, the traits are ordered by their desirability values, from more undesirable to more desirable.

Received December 17, 1990
Accepted March 22, 1991 ■