Selective attention and subjective confidence calibration

JORDAN R. SCHOENHERR, CRAIG LETH-STEENSEN, AND WILLIAM M. PETRUSIC
Carleton University, Ottawa, Ontario, Canada

In the present experiments, failures of selective visual attention were invoked using the B. A. Eriksen and C. W. Eriksen (1974) flanker task. On each trial, a three-letter stimulus array was flashed briefly, followed by a mask. The identity of the two flanking letters was response congruent, neutral, or incongruent with the identity of the middle target letter. On half of the trials, confidence ratings were obtained after each response. In the first three experiments, participants were highly overconfident in the accuracy of their responding to incongruent flanker stimulus arrays. In a final experiment, presenting a prestimulus target location cue greatly reduced both selective attention failure and overconfidence. The findings demonstrate that participants are often unaware of such selective attention failures and provide support for the notion that, in these cases, decisional processing is driven largely by the identities of the incongruent flankers. In addition, responding was invariably slower and sometimes more accurate when confidence was required than when it was not required, demonstrating that the need to provide posttrial confidence reports can affect decisional processing. Moreover, there was some evidence that the presence of neutral contextual flanking information can slow responding, suggesting that such nondiagnostic information can, indeed, contribute to decisional processing.

The ability to actively select some aspects of the environment for further processing while attempting to ignore others has now been widely researched by cognitive psychologists. This work has yielded the robust finding that failures of *selective attention* occur when irrelevant distracting stimuli or stimulus dimensions are presented along with target stimuli. It seems that irrelevant stimulus aspects often cannot be ignored and can detrimentally affect subsequent responding. Typically, such performance decrements are measured in terms of changes in measures of either response time (RT) or response accuracy. An interesting open question, though, is the extent to which failures of selective attention become manifest in measures of *metacognitive* awareness, particularly with respect to retrospective reports of *subjective confidence*. Everyone is familiar with cases in which either themselves or others have mistakenly seen or heard something, due to selective attention failure, yet stubbornly have refused to consider that their perception might not have been correct. Such anecdotal evidence suggests that when selective attention fails, fairly drastic mismatches should likely occur between individuals' confidence about the accuracy of their perceptual performances and their actual performance accuracy—that is, that any subsequent confidence judgments will be quite poorly calibrated. In this study, a full examination of this notion will be provided within a controlled experimental setting.

To date, there has been a large body of psychological research devoted to the conditions that lead to the miscali-

bration of confidence judgments, particularly with respect to when people are likely to be overconfident (Erev, Wallsten, & Budescu, 1994; Fischhoff, Slovic, & Lichtenstein, 1977; Gigerenzer, Hoffrage, & Kleinbölting, 1991; Griffin & Tversky, 1992; Juslin, Winman, & Olsson, 2000; Soll, 1996). To our knowledge, though, none of this work has examined the role that selective attention might play in such miscalibration. Hence, both the extent to which and the conditions under which miscalibration of confidence occurs when selective attention fails would serve to add to this body of knowledge and also to inform theories of the confidence process. Such issues could also be regarded as rather important in a practical sense, because there is now beginning to be some evidence that confidence assessments can play an important role in determining subsequent actions (e.g., Dholakia & Bagozzi, 2002). In addition, it is of theoretical importance to consider the degree to which the confidence results themselves can provide converging information about the nature of the selective attentional processing that must be occurring within a selective attention paradigm, such as the one that will be of interest here (viz., the flanker task).

Selective Attention

One paradigm that has been widely used to observe constraints on selective visual attention processing involves B. A. Eriksen and C. W. Eriksen's (1974) *flanker task*. The development of this task, which makes use of a single target stimulus surrounded by flanking stimuli, arose out

of earlier research investigating the effects of locating a target embedded within a field of distractors (e.g., C. W. Eriksen & Hoffman, 1972, 1973). The results of that work were suggestive of a minimal visual attentional channel size that can be focused on a target, with the remainder being filled by any proximal stimuli. B. A. Eriksen and Eriksen modified the target search task, removing the need to search the display for the target by situating it in the middle—hence, allowing selective attention mechanisms solely to be examined. The critical feature of their task was, then, whether or not identification of the flanking stimuli provided information that could be passed on to later information-processing stages.

B. A. Eriksen and Eriksen (1974) found that responses to trials involving congruent flanker stimuli (i.e., ones that were either physically identical to the target or from an alternative stimulus set but were still associated with the same manual response as was the target) had the shortest RTs, whereas responses to the incongruent flanker stimuli (i.e., ones that were associated with a manual response opposite to that for the target) had the longest RTs. The RTs for two conditions in which the flankers were not associated with any actual response were roughly the average of the congruent and incongruent flanker condition RTs. These effects were also mirrored in the accuracy data and were greatly diminished by increasing the spacing between the flankers and target up to 1° of visual angle. All of these findings were consistent with the idea that selective attention failed to tease apart the spatial locations of flanker and target stimuli, so that the identities of the flankers were often processed to the extent to which they represented potential alternative responses.

By now, it is generally accepted that flanker effects arise at the response generation stage due to either competition or facilitation (for incongruent or congruent flankers, respectively) between response activations associated with both the target and flanker stimuli (Heitz & Engle, 2007; Lavie, 1995; Lavie & de Fockert, 2003; Ridderinkhof, 2002a, 2002b; Sanders & Lamers, 2002). Quite convincing psychophysiological evidence for the presence of such response competition was first provided by Coles, Gratton, Bashore, Eriksen, and Donchin (1985), who used muscular EMG and subthreshold dynamometer squeezing responses to show that both correct and incorrect response channels are much more likely to be concurrently activated when flanker stimuli are incongruent (see also Rosler & Finger, 1993).

Confidence

Confidence calibration. The precision with which individuals can judge the accuracy of their own performance is known as confidence calibration and provides insight into their metacognitive awareness of the quality of that performance. In quantitative terms, confidence *calibration* is the degree to which subjective probability assessments accurately represent the actual relative frequencies of events (e.g., being correct or not). Two other confidence indices are *over-/underconfidence*, which indicates the direction and extent to which an individual's overall subjective performance estimates deviate from overall perfect

calibration (Lichtenstein & Fischhoff, 1977), and *resolution*, which indicates the extent to which an individual can distinguish between an event's occurrence and nonoccurrence (Murphy, 1973).

In much experimental work on confidence in decision making, a rating scale is used to obtain retrospective confidence reports in a local fashion after each decision has been made. Customarily, the ratings options on such a scale run through the possibilities from *guess* to *completely certain*, although, in order to obtain actual calibration measures, it is more useful to have individuals rate their confidence on a scale of 50%–100% *certain*. In the latter case, as will be detailed later, the deviations of the observed accuracies from the numerical points on this scale can be used to derive indices for each of the three aforementioned calibration measures. A number of conceptual and methodological issues surrounding the use of this kind of approach have been astutely reviewed by Keren (1991).

Overconfidence. Although the calibration and resolution indices are often reported in confidence calibration research, the over-/underconfidence (bias) index is reported most frequently by far. Indeed, an examination of the literature provides overwhelming evidence of overconfidence in a wide variety of both laboratory and real-world (e.g., Billet & Qian, 2008; Cesarani, Lichtenstein, Johannesson, & Wallace, 2009; Christensen-Szalanski & Bushyhead, 1981; Oskamp, 1965; Smith & Dumont, 2002; Walter et al., 1998) contexts. Laboratory work on the study of overconfidence typically involves an assessment of participants' general knowledge. A series of almanac-type questions is presented (e.g., "absinthe is a precious stone—True or False"), and following each decision, participants rate the confidence that they made the correct choice. Fischhoff et al. (1977) showed that participants were not nearly 100% correct when they indicated they were 100% confident, thus providing clear evidence of overconfidence within the cognitive-conceptual domain. Lichtenstein and Fischhoff (1977) extended this work by showing that overconfidence was especially evident for their harder subset of the test items, whereas underconfidence occurred for their relatively easier subset of the items. Subsequently, this finding was dubbed the *hard-easy effect*, and it turns out to be quite ubiquitous (see Juslin et al., 2000).

The background of overconfidence in the conceptual domain set the stage for extensions of calibration work to the perceptual domain. For example, Baranski and Petrusic (1994) showed, in a binary visual proximity discrimination task, that when the comparative judgments were relatively easy, participants' judgments of the confidence that they had rendered a correct judgment underestimated the probability that they were in fact correct (i.e., they were underconfident). On the other hand, when difficult comparisons were required, participants were highly overconfident. Thus, the hard-easy effect evident in conceptually based judgments is indeed paralleled in the perceptual domain. Subsequently, Baranski and Petrusic (1995, 1999) and Petrusic and Baranski (1997) replicated and extended these findings (see also Bar-tal, Sarid, & Kishon-Rabin,

2001; Kvidera & Koutstaal, 2008). Moreover, Lucas and Petrusic (2009), in a visual length comparison task, recently replicated an additional finding that overconfidence is greater under speed stress than under accuracy stress, although, in each case, overconfidence increases with increases in decisional difficulty.

Theories of the confidence process. One very important initial class of theories comprises the *signal detection*, or *strength-based*, theories (Ferrell & McGoey, 1980; Wallsten & Gonzalez-Vallejo, 1994; for a discussion of the use of such models within the recognition memory domain, see also Busey, Tunnicliff, Loftus, & Loftus, 2000; Van Zandt, 2000), which assume that decisional information involving the strength of evidence for one of two possible stimulus events (e.g., signal or noise, each of which is associated with a distribution of possible strength values) is directly scaled to derive confidence ratings. One important limitation of the signal detection view, however, is that it does not provide direct accounts for the relationships that have been observed between primary decision RTs and confidence levels—namely, that RTs typically decrease as the level of confidence increases and, more important, that they can also differ across decisional difficulty conditions within each confidence level (for a full discussion of this issue, see Baranski & Petrusic, 1994).

One fruitful approach to extending the notions espoused by the signal-detection-based theories has been to assume that decisional strength information is sampled sequentially. One particularly powerful version of such a sequential-sampling model has recently been developed by Van Zandt (2000; Merkle & Van Zandt, 2006). In this model, stimulus information is sampled over time in discrete packets. For the two-choice case, each packet advances the selection of one of the possible responses (in a probabilistic fashion that is based on corresponding sampled strength values) and is accrued in terms of a count in one of two separate counters. Responses can be triggered whenever the accrued total of counts in one of the counters reaches a certain criterion level (cf. Laberge, 1962). It is assumed that, after this decision process has finished on each trial, the confidence judgment is based on the difference between the winning and losing counter totals (i.e., the *balance-of-evidence* hypothesis; for a demonstration of how this difference can be scaled into a confidence judgment, see Merkle & Van Zandt, 2006, and note that this notion was first proposed by Vickers, 1979, within the context of a model for which strength values are assumed to be accumulated directly). When the difference between resultant counter totals is large, confidence will be correspondingly higher.

A closely related model is the *slow-and-fast-guessing* model, which was first introduced by Petrusic (1992) and then extended to the realm of confidence judgments by Petrusic and Baranski (1997; Baranski & Petrusic, 1996, 1998, 2003). The key facet of this model is that it represents an explicit attempt to accommodate occasions in the evidence accrual process where more neutral information might, in fact, be accrued that is related to neither response. Hence, the slow-and-fast-guessing model assumes the presence of an additional counter for the event

uncertain or *doubtful*, and when it reaches criterion, a guess response is activated. In these terms, confidence can then be derived as a function of the amount of information that has been accumulated for the eventual winning response, relative to that which has been accumulated for all other kinds of competing information (i.e., that favoring both the alternative losing response and also nondiagnostic information).

A second class of theories consists of the *cue-based* theories (Gigerenzer et al., 1991; Koriatic, 1993, 1995). Although such theories have mainly been applied to confidence reports obtained in both general knowledge (Gigerenzer et al., 1991) and memory (Koriatic, 1993, 1995) tasks, there is no reason why they could not be extended to account for confidence reports obtained in perceptual judgment tasks. According to Gigerenzer et al.'s view, the decision maker constructs a mental model of the decision task that allows either for a direct 100% certain response or for the retrieval and examination of contextually relevant informational cues that are only probabilistically associated with the actual correct response. Such cues are assumed to be retrieved sequentially according to their cue validity for the decision problem, and confidence is assumed to be directly related to cue validity of the cue that is used to make the response.

According to Koriatic's (1993, 1995) view, people use whatever kind of information is available to them when making confidence judgments. Such information can include the utilization of either intrinsic cues associated with the stimuli (e.g., how quickly they can be encoded) or extrinsic cues associated with the task situation itself (e.g., an individual's perceived expertise). As has been discussed by Busey et al. (2000), the utilization of both cue- and strength-based information can occur in unison, with cue-based information either inflating or deflating the amount of confidence derived from strength-based information. In addition, Busey et al. regarded the cue utilization process to be more *analytical* than the strength utilization process, because it necessarily involves an overt, heuristical consideration of how the decision process might be impacted by the presence of the cue.

The locus of confidence processing. Two important issues surrounding the rendering of confidence that have received quite a bit of research attention from the third author of this article concern the point within the decision process at which confidence judgments are made and, more generally, the nature of the architecture of the decision–confidence relation. Sequential-sampling models invariably assume that confidence is derived postdecisionally from information accumulated over the course of the decision process and that, once a decision has been reached, a mere review of this information can produce a confidence report (for a full discussion of this issue, see Baranski & Petrusic, 1998). Similarly, both the signal-detection-based and the mental model, cue-validity-based theories assume that confidence is derived at the same time as the decision. Such assumptions, however, have not been supported by studies of the time needed to provide the confidence judgments themselves (i.e., confidence RT), where speed and accuracy are differentially stressed.

In accuracy emphasis conditions, confidence processes require little additional time after the response is rendered. In speeded conditions, though, participants are seemingly forced to compute confidence after the decision, causing substantial increases in the time to provide confidence, relative to the accuracy emphasis conditions (Baranski & Petrusic, 1998). Moreover, there is now further key evidence that the primary decision process itself can be affected by the subsequent provision of confidence. This evidence has been obtained by comparing primary decision task performance under conditions in which participants give subsequent confidence judgments and conditions in which no such judgments are required. Such research has clearly shown that primary decision RTs are lengthened when confidence is required (Baranski & Petrusic, 2001; Petrusic & Baranski, 2003).

The Present Study

The main focus of the present work will essentially be to answer the following question: How is confidence affected by failures of selective visual attention? The selective attention paradigm used herein will be the B. A. Eriksen and Eriksen (1974) flanker task. Standard RT and accuracy performance measures, in addition to confidence ratings, will be obtained under varying conditions of flanker congruency. Of major interest will be analyses involving confidence calibration measures. In general, we expect failures to selectively attend to the target on trials with incongruent flanker stimuli to lead to overconfidence with respect to correctly identifying the targets. As we will then argue, we believe that such a result is consistent with the currently accepted theoretical views that we have reviewed concerning both the manner in which the flanker interference effect is assumed to arise and the manner in which confidence reports are assumed to be generated.

The only previously conducted research that is at all related to this kind of question was performed by Keren (1988). In Keren's (1988) experiment of interest here (Experiment 2), participants viewed two letters for about 50–85 msec and then reported the identity of a target letter that was indicated by a poststimulus position cue. The letters were presented simultaneously in the middle of a 3×4 matrix that contained a premask of dollar signs initially and a postmask of number signs (in addition to a position cue arrow). Similar to the B. A. Eriksen and Eriksen (1974) flanker task, the uncued letter could be either the same as or different from the cued letter. Keren (1988) found that accuracy performance was substantially degraded when the target letter was flanked by itself rather than by another letter (i.e., a *repeated letter inferiority effect* of 61% vs. 81% correct, respectively), which, interestingly, was opposite to the effect that would have been expected on the basis of numerous previous results obtained using B. A. Eriksen and Eriksen's task. Nevertheless, this type of finding is actually a very robust one, which is now more commonly known as the *negative flanker effect* (Rouder & King, 2003; van Leeuwen & Lachmann, 2004). With respect to confidence measures, though, Keren's (1988) participants were quite overconfident in the repeated-letter condition and quite undercon-

fident otherwise (i.e., the mean confidence ratings were 70% and 74%, respectively).

However, there are a number of important aspects of Keren's (1988) findings that indicate that additional research on this question is needed. First, because they involved negative flanker accuracy effects, the exact nature and locus of the cognitive processes that gave rise to them is still not at all clear (Keren, 1988; Rouder & King, 2003; van Leeuwen & Lachmann, 2004; see also Klapp & Hinkley, 2002). Interestingly, Rouder and King have made the case that the presence of such effects implies that flanker stimuli must, indeed, affect both the perceptual encoding and response generation stages of processing in these kinds of tasks (although the former effect is likely to become more pronounced only when perception of the stimuli is made difficult). Nonetheless, as also was discussed by those researchers, regardless of the direction of the flanker effects in any particular situation, they still serve to implicate failures of selective attention processing, because they indicate that flanker stimuli somehow affect the processing of the target.

Second, given the overall observed differences in accuracy between the flanker congruency conditions in Keren's (1988) Experiment 2, the phenomenon of overconfidence in the repeated-letter condition could simply represent a manifestation of the *hard–easy*, or the *calibration difficulty*, effect—namely that, as was mentioned earlier, participants tend to be overconfident when engaged in more difficult decisions (as operationalized by low accuracy levels) but underconfident when decisions are easy (Gigerenzer et al., 1991; Griffin & Tversky, 1992; Keren, 1991; Lichtenstein & Fischhoff, 1977). As was demonstrated by Griffin and Tversky, such an effect can occur simply because the overall decisional difficulty differences are not taken into account when deriving confidence (e.g., the fact that similar subjective strength values might not actually be as diagnostic in the a priori more difficult decisional case than in the easier one is not considered by the decision maker). Alternatively, hard–easy effects can also arise simply because participants are anchoring and adjusting their confidence judgments to reflect what they perceive as the overall intermediate level of difficulty of the tasks that they are currently engaged in (Keren, 1991).

Third, given that the target stimulus was only postcued, unlike for the B. A. Eriksen and Eriksen (1974) task, its location within the array was not known in advance. Hence, it seems likely that the participants in Keren's (1988) Experiment 2 would always have attempted to initially process both of the letter stimuli on each trial. In this case, although one of the letter stimuli was eventually to be ignored, this task cannot then technically be regarded as a pure selective attention task.

Generally then, it seems that whereas Keren's (1988) findings do represent an important contribution with respect to how it is that the identity of irrelevant stimulus information included in the decisional process can affect confidence, Experiments 1 and 2 of the present study will add to that work by making an explicit attempt to examine the effects on confidence of erroneously attend-

ing to information that is only tangential to the decision task allotted to the participant and also by attempting to equate the overall accuracy obtained for congruent and incongruent flanker stimulus conditions so that these effects can be examined in a way that is not then potentially confounded by hard–easy effects. In Experiment 3, we attempted to determine whether the modifications made to the experimental protocol to equate accuracy levels across flanker congruency conditions in Experiment 2 might be providing explicit cues that could then be inducing any observed overconfidence. Finally, in Experiment 4, we attempted to determine whether a modification designed to enhance selective attention to the target would, in turn, serve to attenuate any tendencies toward overconfidence for incongruent flanker stimuli.

It is important to note, as well, that a couple of other issues were also addressed in this study. The first one involved an attempt to replicate the aforementioned findings of Baranski and Petrusic (2001) and Petrusic and Baranski (2003) regarding the effect of rendering confidence on performance of the primary decision process. To this end, participants always performed both a block of trials in which confidence judgments were required and an identical block of trials in which confidence judgments were not required. Second, because flanker congruity manipulations directly entail manipulating the extraneous contextual nature of the information presented to participants during the decision task, the full set of RT, accuracy, and confidence performance measures obtained here across flanker congruity conditions should also help to shed some light on the nature of the decisional processes used. More specifically, it will be of interest to determine the role that irrelevant neutral flanking information plays in the decision process. Because such information could be regarded as truly being nondiagnostic, evidence for the explicit accumulation of such information could provide some support for the uncertain accumulator assumption of Petrusic's (1992; Baranski & Petrusic, 1996, 1998, 2003; Petrusic & Baranski, 1997) slow-and-fast-guessing model.

EXPERIMENT 1

Method

Participants. Twenty undergraduates participated in a single 45-min session for course credit at Carleton University. All the participants had normal or corrected-to-normal vision.

Stimuli and Apparatus. The stimuli consisted of three capital letters, a target in the center of an array, with two identical flankers on either side. Two sets of targets and flankers were used: F, P, K, T and H, N, L, M (the latter set was assumed, *a priori*, to be more difficult than the former). The first two letters in each set were used as both targets and flankers, and the last two were used only as neutral nontarget flankers (which were presented equally often with each of their corresponding target stimuli). These stimuli were presented side by side in the center of a computer monitor in standard DOS font size. The actual width and height of each letter on the screen was 3×5 mm, with a 1-mm spacing between letters. Following a short blank period after the initial presentation of the stimuli, a pattern mask (#####) was presented in which the middle three #s were situated in the same location as the initial three-letter array. A 486

computer using Mel 2.0 experimental software randomly presented the stimuli within each block.

Procedure. The participants were instructed to identify the middle target letter in a rapidly presented display of three letters and were shown all possible stimulus combinations in advance of the experiment. A short initial practice phase of 20 trials familiarized the participants with the procedure. Two experimental blocks of 144 trials (24 trials per flanker congruency condition for each stimulus set) were performed by each participant, one block in which confidence was also required and one block in which the primary decision was the sole requirement (the order in which each participant performed each of these two blocks was counterbalanced across participants).

Each trial began with the cue "Ready" presented in the center of the top portion of the screen for 2,500 msec. After a blank foreperiod of 1,000 msec, the stimulus was presented in the horizontal and vertical center of the screen. Because the interest was in confidence processing under conditions in which a substantial amount of errors would be made, the stimuli were presented for only a very short duration of 17 msec. (Note that event times in MEL 2.0 can be set only in units of 16.67 msec, due to constraints associated with the refresh rate of the monitor.) Their presentation was followed by a blank screen of 50, 67, or 83 msec for the congruent, neutral, and incongruent flanker stimuli, respectively (these times differed because some pilot work indicated that responding would otherwise have been either much too accurate for the congruent flanker stimuli or much too inaccurate for the incongruent flanker stimuli). The pattern mask followed and remained on the screen until the response had been made.

Using the index fingers of each hand, the participants selected the appropriate response key (either "z" or "/") when they had determined the identity of the target. Two targets, one from each different stimulus set (F, H and P, N, respectively), were assigned to each response key. If the block of trials required confidence judgments, once the primary decision had been rendered, a prompt to indicate the level of confidence that the participants had in their response would follow. Confidence was indicated by pressing number keys (on the top row of the computer keyboard) labeled from 50% to 100% confidence. For the no-confidence block, a corresponding prompt appeared that simply instructed the participants to press the space bar to continue to the next trial.

Results

Two separate 2 (confidence block) $\times 3$ (flanker congruency) $\times 2$ (stimulus set) $\times 2$ (confidence block order) ANOVAs, with the latter factor being only between participants, were performed on the mean participant decisional RTs and proportions correct (PCs). For these analyses, PCs were arcsine transformed, and RTs beyond three standard deviations of the overall mean RT of each participant within each block, as well as those below 200 msec, were excluded (2.14%) before computing each participant's mean RT for each cell of the design. For additional analyses involving mean error RTs, any cells with missing values (8 out of 240 for Experiment 1) were replaced by assuming additivity between the cell and participant effects (i.e., by using the corresponding participant mean + cell mean – grand mean values). All significant results for the key confidence block (i.e., confidence given or not given) and flanker congruency factors are reported with their Greenhouse–Geisser adjusted levels of significance. For the sake of brevity, all of the results of the significance tests involving either the stimulus set or the confidence block order factors have been omitted, except for the cases in which they significantly interacted with the confidence block factor.

Table 1
Average Reaction Times (RTs, in Milliseconds) and
Proportions Correct (PCs) in Experiment 1 by
Confidence Block, Flanker Congruency, and Stimulus Set

Confidence Block	Stimulus Set	Flanker Congruency					
		Congruent		Neutral		Incongruent	
		RT	PC	RT	PC	RT	PC
CON	1	1,641	.783	1,662	.825	1,527	.735
	2	1,852	.738	1,922	.644	1,812	.475
	<i>M</i>	1,746	.760	1,792	.734	1,670	.605
NCON	1	1,280	.783	1,295	.802	1,266	.713
	2	1,467	.721	1,594	.654	1,395	.513
	<i>M</i>	1,373	.752	1,444	.728	1,330	.613

Note—CON, confidence reporting; NCON, no confidence reporting.

In the RT analyses, the main effect of confidence block was significant [$F(1,18) = 18.64, p < .001, \eta_p^2 = .509$], so that when confidence judgments were required, the participants showed dramatic increases in RT (1,736 vs. 1,383 msec, respectively; see also Table 1). The flanker congruency effect was not significant for RT [$F(2,36) = 2.62, p > .106, \eta_p^2 = .127$, where the overall mean RTs were 1,560, 1,618, and 1,500 msec for the congruent, neutral, and incongruent flanker conditions, respectively]. Analogous ANOVA results were obtained for a separate analysis of the mean correct RTs (for which the means were 1,513, 1,562, and 1,510 msec for the congruent, neutral, and incongruent flanker conditions, respectively). However, a separate analysis of the mean error RTs also revealed a significant flanker congruency effect [$F(2,36) = 7.50, p < .002, \eta_p^2 = .294$, where the means were now 1,778, 1,868, and 1,577 msec for the congruent, neutral, and incongruent flanker conditions, respectively].

In the PC analyses, the main effect of confidence block was not significant [$F(1,18) = 0.13, p > .724, \eta_p^2 = .007$; see Table 1]. Flanker congruency was significant for PC [$F(2,36) = 8.17, p < .005, \eta_p^2 = .312$], in that response

accuracy was comparatively degraded for the incongruent flanker stimulus trials.

Four separate 3 (flanker congruency) $\times$ 2 (stimulus set) ANOVAs were also performed on the individual mean confidence measures and the calibration indices of over-/underconfidence, calibration, and resolution. Group confidence calibration curves derived using averaged PCs at each confidence level are given in Figure 1. Mean confidence level was significantly affected by flanker congruency [$F(2,38) = 4.75, p < .023, \eta_p^2 = .200$; see Table 2]. Over-/underconfidence, which was computed as the difference between mean confidence divided by 100 and the corresponding PC measure, also exhibited significant effects of flanker congruency [$F(2,38) = 13.25, p < .001, \eta_p^2 = .411$].

Measures of calibration were computed as the sum of the squared difference between each confidence level (divided by 100) and the observed PC achieved within each of those respective reported confidence levels, where each squared difference was then weighted by the relative frequencies for each confidence level (e.g., Baranski & Petrusic, 1994; Petrusic & Baranski, 1997).

Figure 1. Group confidence calibration curves derived by plotting the averaged proportion correct (PC) within each confidence-level category for the responding to each flanker congruency condition and each stimulus set in Experiment 1. The percentages for frequency of usage of the confidence categories associated with each of the plotted points are indicated by the accompanying inset values.

Table 2
Average Calibration Indices in Experiment 1
by Stimulus Set and Flanker Congruency

	Congruency	<i>P</i> (correct)	<i>M</i> _{conf}	O/U	Cal	Res
Stimulus Set 1	Congruent	.784	74.5	−0.039	0.045	0.040
	Neutral	.828	78.9	−0.038	0.052	0.053
	Incongruent	.739	78.8	0.049	0.061	0.046
Stimulus Set 2	Congruent	.750	69.5	−0.055	0.059	0.034
	Neutral	.651	71.4	0.064	0.060	0.052
	Incongruent	.473	71.3	0.240	0.172	0.044
<i>M</i>		.704	74.1	0.037	0.075	0.045

Note—Proportion correct measures are slightly different here than in Table 1 because they correspond only to responses where the associated confidence judgments given were within 50%–100%. *M*_{conf}, mean confidence; O/U, over-/underconfidence; Cal, calibration; Res, resolution.

This measure essentially provides an index of how close the observed calibration curves are to the curve that represents perfect calibration (i.e., the solid diagonal curves in Figure 1), with smaller values representing better calibration. The ANOVA results performed on these measures obtained from the individual calibration curves for each of the participants indicated that calibration also differed significantly with flanker congruency [$F(2,38) = 10.11, p < .003, \eta_p^2 = .347$], due mainly to the fact that calibration was much poorer in the incongruent flanker condition.

Individual measures of resolution for each participant in each condition were computed as the sum of squared difference between the participant's overall PC observed for that condition and the observed accuracy achieved within each of the respective reported confidence levels, where each squared difference was also then weighted by the relative frequencies for each confidence level (e.g., Baranski & Petrusic, 1994; Petrusic & Baranski, 1997). This measure essentially provides an index of the extent to which the different confidence categories are being used in a differentiating fashion with respect to the corresponding accuracies (i.e., not all responses in one or two categories and also not a flat confidence calibration curve), with larger values now representing better resolution. In contrast to the other confidence-related results, no significant factor effects were present for the resolution measures.

Discussion

The key result from this experiment was the presence of substantial increases in overconfidence for the incongruent flanker conditions, relative to the congruent and neutral flanker conditions. This finding was also accompanied by a corresponding degradation of confidence calibration in the incongruent flanker conditions.

A key secondary result was the fact that decisional RT was much longer in the confidence blocks than in the no-confidence blocks of this experiment. As such, this result represents a clear replication of Baranski and Petrusic's (2001) and Petrusic and Baranski's (2003) findings regarding the effect of rendering confidence on the primary decision process.

EXPERIMENT 2

One aspect of the confidence calibration results of Experiment 1, however, is that, due to the overall differences in accuracy levels across flanker congruency conditions, these results (like Keren's [1988]) are still potentially confounded by hard–easy calibration difficulty effects. This issue was addressed in Experiment 2 by increasing the length of the blank period between the stimuli and the mask in the incongruent flanker condition so as to attempt to minimize differences in accuracy across flanker congruency conditions. Such a manipulation should help to selectively isolate any effects on confidence processing that are due solely to differences in the types of flankers used in each condition.

Method

Twenty new undergraduates participated in a single 45-min session. The stimuli and procedures used were identical to those in Experiment 1, with the exception that the poststimulus blank periods before the mask on the congruent, neutral, and incongruent flanker stimulus trials were now set to be 50, 67, and 117 msec, respectively. These values were chosen because the results of Experiment 1 showed that accuracy for the congruent and neutral flanker conditions was quite comparable when the poststimulus blank periods were set to 50 and 67 msec, respectively. Furthermore, some additional pilot work indicated that increasing the length of this period to 117 msec for the incongruent flanker condition would make the accuracy for that condition more comparable to that for the other two conditions.

Results

The details of the statistical analyses of the decisional RTs and PCs were exactly the same as those in the previous experiment (although 2.50% of the RT data were excluded here). In the RT analyses, the main effect of confidence block was again significant [$F(1,18) = 13.69, p < .002, \eta_p^2 = .432$], so that when confidence judgments were required, the participants showed dramatic increases in RT (1,438 vs. 1,075 msec, respectively; see Table 3). However, in this experiment, confidence block order interacted significantly with confidence block [$F(1,18) = 9.74, p < .006, \eta_p^2 = .351$], so that the RT difference when confidence was and was not required was much smaller when the confidence judgment block was performed second

Table 3
Average Reaction Times (RTs, in Milliseconds) and
Proportions Correct (PCs) in Experiment 2 by
Confidence Block, Flanker Congruency, and Stimulus Set

Confidence Block	Stimulus Set	Flanker Congruency					
		Congruent		Neutral		Incongruent	
		RT	PC	RT	PC	RT	PC
CON	1	1,328	.804	1,426	.821	1,249	.813
	2	1,511	.694	1,666	.706	1,447	.625
	<i>M</i>	1,420	.749	1,546	.764	1,348	.719
NCON	1	1,018	.773	1,110	.808	944	.752
	2	1,124	.694	1,180	.654	1,075	.565
	<i>M</i>	1,071	.733	1,145	.731	1,009	.658

Note—CON, confidence reporting; NCON, no confidence reporting.

(56 msec) than when it was performed first (670 msec). In addition, for this experiment, the main effect of flanker congruency was now significant for RT [$F(2,36) = 13.83$, $p < .001$, $\eta_p^2 = .434$, where the overall mean RTs were 1,246, 1,346, and 1,179 msec for the congruent, neutral, and incongruent flanker conditions, respectively]. Analogous ANOVA results were obtained for separate analyses of both the mean correct and mean error RTs (for which, mean correct RTs were 1,199, 1,324, and 1,164 msec, and mean error RTs were 1,465, 1,589, and 1,263 msec, for the congruent, neutral, and incongruent flanker conditions, respectively, where, in the latter analysis, there were now 29 out of 240 missing cells, 13 of which were in the neutral flanker condition).

In the PC analyses, the main effect of confidence block was now significant [$F(1,18) = 7.03$, $p < .016$, $\eta_p^2 = .281$], so that when confidence judgments were required, the participants were more accurate (.744 vs. .708, respectively; see Table 3). Flanker congruency was now not significant for PC [$F(2,36) = 1.50$, $p > .239$, $\eta_p^2 = .077$]. Importantly, mean PCs were very similar across flanker congruency conditions in this experiment for the more

accurately responded to F, P Stimulus Set 1, indicating that, at least for this stimulus set, stimulus presentation time manipulations were successful in eliminating accuracy differences.

The statistical analyses performed on the individual mean confidence and the three calibration indices were also the same as those in the previous experiment (group confidence calibration curves are given in Figure 2). Mean confidence level was again significantly affected by flanker congruency [$F(2,38) = 9.45$, $p < .001$, $\eta_p^2 = .332$; see Table 4]. Importantly, over-/underconfidence also still exhibited significant effects of flanker congruency [$F(2,38) = 6.08$, $p < .013$, $\eta_p^2 = .243$]. Overconfidence was essentially twice as high for incongruent flanker stimuli. Calibration also differed with flanker congruency (worse for incongruent flanker stimuli) in a marginally significant fashion [$F(2,38) = 3.35$, $p < .069$, $\eta_p^2 = .150$].

Discussion

The key result from this experiment was, once more, the finding of sizably enhanced overconfidence in the in-

Figure 2. Group confidence calibration curves derived by plotting the averaged proportion correct (PC) within each confidence-level category for the responding to each flanker congruency condition and each stimulus set in Experiment 2. The percentages for frequency of usage of the confidence categories associated with each of the plotted points are indicated by the accompanying inset values.

Table 4
Average Calibration Indices in Experiment 2
by Stimulus Set and Flanker Congruency

	Congruency	<i>P</i> (correct)	<i>M</i> _{conf}	O/U	Cal	Res
Stimulus Set 1	Congruent	.809	87.4	0.065	0.060	0.034
	Neutral	.825	87.1	0.046	0.047	0.038
	Incongruent	.814	92.5	0.111	0.072	0.013
Stimulus Set 2	Congruent	.714	78.8	0.074	0.069	0.056
	Neutral	.716	79.2	0.076	0.074	0.030
	Incongruent	.626	83.7	0.211	0.138	0.041
<i>M</i>		.751	84.8	0.097	0.077	0.035

Note—Proportion correct measures are slightly different here than in Table 3 because they correspond only to responses where the associated confidence judgments given were within 50%–100%. *M*_{conf}, mean confidence; O/U, over-/underconfidence; Cal, calibration; Res, resolution.

congruent flanker condition, relative to the congruent and neutral flanker conditions, a finding that was again also accompanied by a corresponding degradation of confidence calibration in the incongruent flanker conditions. The fact that enhanced overconfidence for incongruent flanker conditions occurred even in the case in which accuracy was closely equated across flanker congruency conditions (i.e., for Stimulus Set 1) permits the conclusion that it does not simply represent a manifestation of the hard–easy effect but, instead, indicates that failures of selective attention can have real and quite dramatic effects on confidence.

In addition, RT was once more much longer in the confidence blocks than in the no-confidence blocks in this experiment. Unlike in Experiment 1, the size of the RT confidence rendering effect was moderated (but not reversed) by increased practice with the letter identification task, in that it now differed quite substantially depending on the confidence block ordering. Importantly, an effect of confidence rendering was now also evident in the PC measures, so that the participants were more accurate when subsequently rendering confidence than when not.

Finally, RT now also differed significantly with flanker congruency. Notably, RTs were longest for responses to the neutral flanker stimuli and shortest for responses to the incongruent flanker stimuli. As such, the former finding provides some evidence that the presence of irrelevant and completely noninformative information within a stimulus display can slow responding to target items, even relative to congruent and incongruent flanker situations in which either fully confirming or fully competing irrelevant information is present. Regarding the latter finding, although it does not represent the typical flanker RT effect, note that (1) the mean RT difference between the congruent and incongruent flanker conditions was much smaller for correct than for error responses, (2) poststimulus blank periods were much longer in the incongruent flanker condition, and (3) brief stimulus presentation times followed by a subsequent masking did provide the ideal conditions for the elicitation of negative flanker RT effects (Rouder & King, 2003; in fact, these researchers have postulated that the reason negative flanker effects are not commonly observed is that they are typically overwhelmed by strong response competition effects).

EXPERIMENT 3

One aspect of Experiment 2, however, was that in order to make the accuracy of responding to the incongruent flanker stimuli more equivalent to the accuracy for the other two flanker congruency conditions, the delay time after the presentation of the stimuli and before the presentation of the mask for the incongruent flanker stimuli was increased to about twice that used in the other two conditions. Given this difference, it is possible that the participants could have noticed it and used it as an intrinsic cue to inflate their confidence reports to these stimuli (i.e., the fact that these stimuli seemed slightly more visible might have influenced the participants to believe that they should be more confident in their responses; Koriati, 1993, 1995). This issue was addressed in Experiment 3 by providing accuracy feedback after every trial.

Interestingly, past research has often shown that the use of feedback does not generally seem to have much of an effect on confidence calibration. For example, Keren (1988) found no effect of trial-by-trial feedback on over-/underconfidence in his Experiment 1 (involving both general knowledge of city population sizes and perceptual gap detections). In Baranski and Petrusic (1994), the provision of trial-by-trial accuracy feedback improved resolution for fairly difficult perceptual decisions but did not affect the calibration measure. Petrusic and Baranski (1997) did find some effects of such feedback on overconfidence and calibration, which, however, depended in a fairly complex fashion on the easiness or hardness of the surrounding contextual difficulty.

Small or null effects of trial-by-trial feedback on confidence calibration could be expected if confidence is solely the result of nonanalytical processes involving implicit assessments of strength values, accumulator totals, or cue validities. However, trial-by-trial feedback would be expected to affect confidence reporting if such reports are influenced by explicit, analytical considerations of the decisional environment. Hence, if the participants used the increased delay time prior to masking in the incongruent flanker conditions as an intrinsic cue to substantially raise their confidence reporting in that condition, trial-by-trial feedback should make them aware that this cue itself was not a very valid one. In that case, the amount of overcon-

Table 5
Average Reaction Times (RTs, in Milliseconds) and
Proportions Correct (PCs) in Experiment 3 by Confidence Block,
Flanker Congruency, and Stimulus Set

Confidence Block	Stimulus Set	Flanker Congruency					
		Congruent		Neutral		Incongruent	
		RT	PC	RT	PC	RT	PC
CON	1	1,286	.773	1,272	.802	1,207	.780
	2	1,446	.716	1,454	.636	1,401	.606
	<i>M</i>	1,366	.744	1,363	.719	1,304	.693
NCON	1	968	.743	991	.791	932	.753
	2	1,057	.689	1,112	.632	1,041	.589
	<i>M</i>	1,013	.716	1,051	.711	986	.671

Note—CON, confidence reporting; NCON, no confidence reporting.

fidence in those conditions should be reduced from that observed in Experiment 2.

Method

Twenty new undergraduates participated in a single 45-min session. The stimuli and procedures used were identical to those in Experiment 2, with the exception that after each trial (including the confidence report in the block requiring them), feedback about the accuracy of the response just made was given to the participant. This feedback was presented for 2,000 msec, and then the next trial was initiated.

Results

The details of the statistical analyses of the decisional RTs and PCs were exactly the same as those for the previous experiments (although 1.96% of the RT data were excluded here). In the RT analyses, the main effect of confidence block was again significant [$F(1,18) = 45.84, p < .001, \eta_p^2 = .718$], so that when confidence judgments were required, the participants showed dramatic increases in RT (1,344 vs. 1,017 msec, respectively; see Table 5). As in Experiment 2, confidence block order interacted significantly with confidence block [$F(1,18) = 6.79, p < .018, \eta_p^2 = .274$], so that the RT difference when confidence was and was not required was smaller when the confidence judgment block was performed second (201 msec) than when it was performed first (453 msec). As in Experiment 1, the flanker congruency effect was not significant for RT here [$F(2,36) = 2.19, p > .142, \eta_p^2 = .108$, where the overall mean RTs were 1,189, 1,207, and 1,145 msec for the congruent, neutral, and incongruent flanker conditions, respectively]. Analogous ANOVA results were obtained for separate analyses of both the mean correct and mean error RTs (for which, mean correct RTs were 1,155, 1,173, and 1,132 msec, and mean error RTs were 1,368, 1,402, and 1,322 msec, for the congruent, neutral, and incongruent flanker conditions, respectively, where, in the latter analyses, there were now 13 out of 240 missing cells).

In the PC analyses, as in Experiment 1, the main effect of confidence block was not significant [$F(1,18) = 1.28, p > .273, \eta_p^2 = .066$; see Table 5]. As in Experiment 2, flanker congruency was again not significant for PC [$F(2,36) = 0.77, p > .451, \eta_p^2 = .041$].

The statistical analyses performed on the individual mean confidence and the three calibration indices were also the same as those in the previous experiments (group confidence calibration curves are given in Figure 3). Mean confidence level was again significantly affected by flanker congruency [$F(2,38) = 6.89, p < .004, \eta_p^2 = .266$; see Table 6]. Importantly, over-/underconfidence still exhibited significant effects of flanker congruency [$F(2,38) = 4.42, p < .022, \eta_p^2 = .189$]. Calibration also differed significantly with flanker congruency [$F(2,38) = 5.22, p < .020, \eta_p^2 = .216$].

Discussion

The key result from this experiment was that, once again, significantly more overconfidence and worse calibration were found in the incongruent flanker condition, in comparison with the congruent and neutral flanker conditions, even when accuracy feedback was given to the participants on a trial-by-trial basis. Indeed, for at least the Stimulus Set 1 results, the confidence and accuracy measures obtained in Experiment 3 (top of Table 6) were almost identical to those obtained in Experiment 2 (top of Table 4). These findings indicate that it is unlikely that an explicit consideration of the length of the poststimulus, premask delay time in the incongruent flanker conditions influenced confidence.

In addition, once more, RT was much longer in the confidence blocks than in the no-confidence blocks in this experiment. As in Experiment 2, the size of the RT confidence-rendering effect was moderated by practice effects but was still quite substantial for both confidence block orderings. With respect to the PC measures, the participants were slightly more accurate when subsequently rendering confidence than when not (.719 vs. .699, respectively), but this difference was not significant.

EXPERIMENT 4

Even given the results obtained in Experiment 3, it is important to determine further whether the overconfidence observed in the incongruent flanker conditions was truly due to failures of selective attention. One way to try to do so would be to invoke a manipulation that

Figure 3. Group confidence calibration curves derived by plotting the averaged proportion correct (PC) within each confidence-level category for the responding to each flanker congruency condition and each stimulus set in Experiment 3. The percentages for frequency of usage of the confidence categories associated with each of the plotted points are indicated by the accompanying inset values.

would enhance selective attention to the targets in the incongruent flanker conditions, while leaving all of the premask delay times over flanker congruency conditions the same. Hence, in this last experiment, the location of the target stimuli was explicitly cued in advance. As has clearly been demonstrated by Paquet and Lortie (1990; Paquet & Craig, 1997), such cuing of the target location serves to enhance early selection of the target stimulus and dramatically attenuate flanker congruency effects. If the overconfidence observed previously in the incongruent flanker conditions was, indeed, due to selective attention failure, reducing the likelihood of such failure should also reduce overconfidence. If, however, confidence was influenced mainly by the intrinsic cues provided by the lengths of the premask delay times, fairly substantial overconfidence should still be observed in the incongruent flanker conditions.

Method

Twenty new undergraduates participated in a single 45-min session. The stimuli and procedures used were identical to those in Experiment 2 (i.e., no feedback), with the exception that during the

“Ready” prompt, a plus-sign fixation point was always presented in the exact location corresponding to the middle target stimulus in the upcoming three-letter stimulus array.

Results

The details of the statistical analyses of the decisional RTs and PCs were exactly the same as those in the previous experiments (although 2.01% of the RT data were excluded here). In the RT analyses, the main effect of confidence block was again significant [$F(1,18) = 28.11, p < .001, \eta_p^2 = .610$], so that when confidence judgments were required, the participants showed dramatic increases in RT (1,459 vs. 1,070 msec, respectively; see Table 7). Confidence block order again interacted significantly with confidence block [$F(1,18) = 7.05, p < .016, \eta_p^2 = .282$], so that the RT difference when confidence was and was not required was smaller when the confidence judgment block was performed second (194 msec) than when it was performed first (584 msec). As in Experiment 2, flanker congruency was again significant [$F(2,36) = 11.46, p < .001, \eta_p^2 = .389$], where the overall mean RTs were 1,312, 1,315, and 1,166 msec for the congruent, neutral, and incongruent

Table 6
Average Calibration Indices in Experiment 3
by Stimulus Set and Flanker Congruency

	Congruency	$P(\text{correct})$	M_{conf}	O/U	Cal	Res
Stimulus Set 1	Congruent	.789	84.2	0.053	0.038	0.040
	Neutral	.815	87.6	0.061	0.041	0.019
	Incongruent	.798	91.3	0.115	0.070	0.022
Stimulus Set 2	Congruent	.727	77.8	0.051	0.050	0.032
	Neutral	.640	78.5	0.144	0.059	0.034
	Incongruent	.608	80.8	0.200	0.111	0.036
<i>M</i>		.730	83.4	0.104	0.062	0.031

Note—Proportion correct measures are slightly different here than in Table 5 because they correspond only to responses where the associated confidence judgments given were within 50%–100%. M_{conf} , mean confidence; O/U, over-/underconfidence; Cal, calibration; Res, resolution.

Table 7
Average Reaction Times (RTs, in Milliseconds) and
Proportions Correct (PCs) in Experiment 4 by
Confidence Block, Flanker Congruency, and Stimulus Set

Confidence Block	Stimulus Set	Flanker Congruency					
		Congruent		Neutral		Incongruent	
		RT	PC	RT	PC	RT	PC
CON	1	1,452	.840	1,302	.923	1,148	.937
	2	1,614	.717	1,685	.682	1,551	.678
	<i>M</i>	1,533	.774	1,493	.802	1,350	.807
NCON	1	998	.875	1,046	.909	872	.899
	2	1,191	.671	1,228	.671	1,093	.668
	<i>M</i>	1,094	.773	1,137	.790	982	.784

Note—CON, confidence reporting; NCON, no confidence reporting.

ent flanker conditions, respectively]. Moreover, the two-way stimulus set $\times$ confidence block interaction was now significant here [$F(1,18) = 7.40, p < .014, \eta_p^2 = .291$], in that RTs were longer for the H, N than for the F, P stimulus set, but more so when confidence was required. Analogous ANOVA results were obtained for separate analyses of both the mean correct and mean error RTs (for which, mean correct RTs were 1,254, 1,278, and 1,158 msec, and mean error RTs were 1,551, 1,554, and 1,416 msec, for the congruent, neutral, and incongruent flanker conditions, respectively, where, in the latter analyses, there were now 35 out of 240 missing cells, 18 of which were in the incongruent flanker condition).

In the PC analyses, as in Experiment 2, the main effect of confidence block was again significant [$F(1,18) = 7.34, p < .014, \eta_p^2 = .290$], so that when confidence judgments were required, the participants were more accurate (.795 vs. .782, respectively; see Table 7). Flanker congruency was once more not significant for PC [$F(2,36) = 1.01, p > .354, \eta_p^2 = .053$].

The statistical analyses performed on the individual mean confidence and the three calibration indices were

also the same as those in the previous experiments (group confidence calibration curves are given in Figure 4). Mean confidence level was significantly affected by flanker congruency [$F(2,38) = 7.20, p < .005, \eta_p^2 = .275$; see Table 8]. However, for both over-/underconfidence and calibration, the main effects of flanker congruency were now not significant [$F(2,38) = 0.68, p > .484, \eta_p^2 = .035$, and $F(2,38) = 0.51, p > .591, \eta_p^2 = .026$, respectively]. For resolution, though, the flanker congruency effect was now significant [$F(2,38) = 4.07, p < .039, \eta_p^2 = .176$], in that this measure was poorest in the incongruent flanker conditions.

Discussion

The key result from this experiment was that accuracy was substantially enhanced by the use of the fixation precue and, now, an overall difference in overconfidence across flanker congruency conditions was not observed. In fact, for Stimulus Set 1, very little overconfidence was observed in any of the flanker congruency conditions (note, as well, that the calibration measure in the incongruent flanker condition for Stimulus Set 1 in this

Figure 4. Group confidence calibration curves derived by plotting the averaged proportion correct (PC) within each confidence-level category for the responding to each flanker congruency condition and each stimulus set in Experiment 4. The percentages for frequency of usage of the confidence categories associated with each of the plotted points are indicated by the accompanying inset values.

Table 8
Average Calibration Indices in Experiment 4
by Stimulus Set and Flanker Congruency

	Congruency	<i>P</i> (correct)	<i>M</i> _{conf}	O/U	Cal	Res
Stimulus Set 1	Congruent	.835	86.1	0.026	0.064	0.025
	Neutral	.922	90.9	−0.013	0.025	0.021
	Incongruent	.935	94.7	0.012	0.015	0.007
Stimulus Set 2	Congruent	.715	78.1	0.066	0.069	0.039
	Neutral	.691	81.5	0.124	0.078	0.044
	Incongruent	.676	83.8	0.162	0.094	0.029
<i>M</i>		.796	85.8	0.063	0.057	0.027

Note—Proportion correct measures are slightly different here than in Table 7 because they correspond only to responses where the associated confidence judgments given were within 50%–100%. *M*_{conf}, mean confidence; O/U, over-/underconfidence; Cal, calibration; Res, resolution.

experiment was better than it had been previously). If confidence was influenced here by both stimulus information and intrinsic, premask delay time cues, such a dramatic drop in overconfidence in the incongruent flanker condition from that obtained previously, especially for Stimulus Set 1, should not have been expected (although it is possible that there could have been some ceiling effects on confidence reporting here, given that, to begin with, mean confidence had been fairly high in this condition in the previous two experiments).

In addition, RT was once more much longer in the confidence blocks than in the no-confidence blocks in this experiment. As in Experiment 3, the size of the RT confidence-rendering effect was moderated by practice effects but was still quite substantial for both confidence block orderings. As in Experiment 2, the participants were also significantly more accurate when subsequently rendering confidence than when not.

GENERAL DISCUSSION

The four experiments performed here were clear in showing that individuals are highly overconfident in their decisional performance under conditions in which their actual performance has been compromised by failures of selective attention. The results for Stimulus Set 1 in Experiment 2 indicate that such findings are not due simply to a manifestation of the hard–easy effect, because large overconfidence in the incongruent flanker condition was found even when accuracy was equated across flanker congruency conditions by varying the amount of delay time before the appearance of the mask in each of these conditions. The results of Experiment 3 indicate that such findings are also not likely a consequence of individuals' adjusting their confidence reports on the basis of an analytical consideration of the amount of delay time before masking, because the provision of explicit trial-by-trial accuracy feedback did not reduce the amount of overconfidence for incongruent flanker stimuli from that observed under identical stimulus presentation conditions in Experiment 2. Finally, when the occurrence of selective attentional failure was substantially attenuated through the use of a location precue in Experiment 4, overconfi-

dence in the incongruent flanker conditions was, in turn, highly attenuated.

More specifically, these findings suggest that the nature of the errors being made in the incongruent flanker conditions was qualitatively quite different from that in the other two flanker conditions, in that, for the former case, such errors were associated with higher overall confidence. Indeed, it seems that in the incongruent flanker condition, the participants were often completely unaware that their selective attention had failed. Such findings are consistent with those obtained in previous investigations of the effect of incongruent flanker stimuli on measures of partial response activation at the psychophysiological level. In particular, Coles et al. (1985) showed that about half of the time when an error occurred to their incongruent flanker stimuli (but much less so, proportionally, for errors to their congruent flanker stimuli), no response-related activation (in terms of subthreshold dynamometer squeezing) was present on the side of the correct response. That finding indicates that much of the time, when an incorrect response to incongruent flanker stimuli "won out," it "won big."

In terms of the available theories of the confidence process, it is possible to account for these overconfidence findings within the general framework of either the strength-based or the cue-validity-based theories of the confidence process. Indeed, with respect to the latter, Gigerenzer et al. (1991) explicitly predicted that their theory should be able to account for overconfidence under perceptual judgment conditions in which the stimulus information provided is somewhat misleading because, in these cases, the generated mental models and resulting cue validities will correspond poorly with what is actually occurring in the environment (i.e., the ecological validities).

With respect to the two major strength-based, sequential-sampling theories (i.e., the race model of Van Zandt [2000] and the slow-and-fast-guessing model of Petrusic & Baranski [1997]), they specifically assume that confidence is derived by comparing the final amounts of covertly accumulated evidence for each response, with higher confidence reports being associated with more discrepancy between those amounts (viz., that confidence is based on the resulting balance of evidence). Hence, if the

present findings are viewed within the framework of these models, they essentially dictate that it must often have been the case that when errors occurred in the incongruent flanker conditions, they occurred with large discrepancies between the dominant and nondominant accumulator totals (which, in this case, were actually associated with the incorrect and correct responses, respectively). Such a state of affairs would also be consistent with the fact that error RTs were typically much shorter in the incongruent flanker conditions than in the congruent and neutral flanker conditions (for some related findings, see Christ, Falkenstein, Heuer, & Hohnsbein, 2000).

It would also be highly constraining on the nature of the processing that must have occurred on many of the incongruent flanker condition error trials (at least under the brief stimulus presentation experimental conditions used here)—namely, that the selective attention failures that occurred within such trials often had an all-or-none character, in that the information accumulation process must have been driven almost exclusively by information about the flanker identities (as opposed to a mix of information about the identities of both the target and flankers, because such a within-trial mix would tend to lessen discrepancies between final accumulator totals). With respect to the still currently popular attentional spotlight view of the flanker effect (Heitz & Engle, 2007), such a state of affairs suggests that on a number of error trials in the incongruent flanker condition, the sampling of stimulus evidence within the attentional spotlight was, for the most part, limited to the locations occupied by the flankers. Regardless, this notion is also highly consistent with the findings of Coles et al. (1985) and also suggests, for the first time ever, that it is an awareness of the response activation of the nature examined by those researchers that could actually be driving metacognitive feelings of confidence.

The second major result was that both RT and accuracy performance measures differed across blocks of trials on which confidence judgments were given and blocks of trials on which they were not given. Namely, not only were RTs significantly longer when confidence was required in all four experiments, but also, in two experiments (2 and 4), accuracy was significantly enhanced for the confidence judgment blocks. As such, the former result represents a clear replication of Baranski and Petrusic's (2001) and Petrusic and Baranski's (2003) findings regarding the effect of rendering confidence on the primary decision process, which now has been demonstrated within a perceptual identification, as opposed to a comparative judgment, paradigm. Moreover, this result has been obtained here in a within-participants fashion under conditions in which the order of the confidence and no-confidence blocks was fully counterbalanced across participants. The latter consideration is key in order to ensure that such an effect is not then confounded by practice effects. Although these RT differences were, indeed, moderated by practice in most of these experiments, such a result is to be expected, since any decreases in RTs for the second block of trials would then serve to work against any increases in RTs when that block also requires confi-

dence judgments. It is also important to note that because the participants were asked to press the space bar in place of providing a confidence judgment in the no-confidence blocks of all four experiments, these results are also not confounded by any potential differences in self-pacing or timing between trial responses across the confidence and no-confidence blocks.

The fact that the provision of confidence can affect performance of the primary decision task itself is an important result that is not readily accommodated by all of the currently available theories of the confidence generation process, given that they all assume that confidence is derived directly from the end result of the decision process (either simultaneously or just after the fact). As has been discussed by Baranski and Petrusic (2001) and Petrusic and Baranski (2003), the finding of lengthened primary decision RTs when confidence is required suggests that some portion of the confidence processing could occur concurrently with the primary decision process (e.g., momentary glimpses of confidence are taken). However, the fact that the present results show that accuracy can sometimes also be enhanced in confidence blocks suggests that although the requirement to provide confidence certainly seems to slow the decision process, it could also sharpen it (e.g., possibly because more attention is then devoted to it). On the other hand, it is possible that the present results indicate that responding simply occurs more cautiously when confidence is required, like that which would be expected if participants had slightly different speed versus accuracy sets across the two types of confidence block conditions (note that this still represents an adjustment of the decision process that, perhaps, occurs dynamically in an adaptive fashion on the basis of an ongoing monitoring of that process). Such a notion actually receives some empirical support here from the fact that the RT differences in Experiment 4 for responding to the two different stimulus sets were significantly larger when confidence was required than when it was not (note also that an analogous marginally significant effect occurred in both Experiments 2 and 3). In general, it can be demonstrated (e.g., Poltrock, 1989) that decision-related RT effects are typically more exaggerated under accuracy than under speed sets (an effect that follows naturally from the assumption that participants wait to accrue more evidence under an accuracy set).

A final result was that RTs were significantly lengthened for the responses to neutral flanker stimuli, in comparison with the responses for the congruent and incongruent flanker conditions in at least one experiment (i.e., Experiment 2). The key aspect of the neutral flanker conditions is that the information being provided by the flanker letters is completely nondiagnostic with respect to the response (remember that neutral flanker letters were always presented with each of the targets in their respective stimulus sets equally often). Hence, the finding of longer RTs in the neutral flanker conditions suggests that such information is being examined in some fashion on those trials. (Although, as one reviewer of this work pointed out, a potential alternative explanation is that the neutral flanker

letters could have been subject to less perceptual priming throughout, given that they were presented less often than the target letters.)

Within the framework of the two sequential-sampling models, although it is quite intuitive to consider that such information is being accumulated as uncertain evidence within the corresponding accumulator in the slow-and-fast-guessing model of Petrusic (1992; Baranski & Petrusic, 1996, 1998, 2003; Petrusic & Baranski, 1997), the accumulation of such evidence can also be accommodated within the more standard race model of Van Zandt (2000) by assuming that its accumulation essentially increments the accumulators for each actual response in a chance fashion (which would then also serve to lengthen RTs). However, at least with respect to the results obtained for Experiment 2 (where this neutral flanker RT effect was the most salient), that information does not seem to have had that much of an effect on the actual outcome of the decision process itself, given that neither accuracy nor overall mean confidence for the neutral flanker stimulus conditions seemed to differ very much from that in the congruent flanker stimulus conditions (note that accuracy was actually a bit higher in the neutral flanker than in the congruent flanker condition there, except when Stimulus Set 2 was responded to in the no-confidence block). Together, these results suggest that during selective attention failure to neutral flanker stimuli, the presence of nondiagnostic information in the decision process does not lead to chance incrementing of the correct and incorrect accumulators (which should then also highly degrade both accuracy and confidence) but, instead, is accumulated in a separate, *slow-guess* uncertain accumulator. Slow-guess accumulators in Petrusic's (1992) model are characterized by response criterion thresholds that are higher than the thresholds for the accumulators for each of the two actual responses themselves. Hence, although slow-guess accumulators can lead to guess responses, more accumulation is needed for this to happen than is needed to generate a nonguess response. Such a state of affairs serves to lengthen both correct and error RTs (because more evidence accruals are needed overall) without drastically affecting accuracy. If it is also assumed that the amount of accumulated evidence for the competing response is given more weight than the amount of accumulated uncertain evidence when confidence is determined, no real differences in overall confidence are expected to occur between neutral and congruent flanker conditions.

Nonetheless, it must be noted that both correct and error RTs were actually quite comparable in both the neutral and congruent flanker conditions in the last two experiments. With respect to those results, we would simply like to suggest that perhaps it is the case that uncertain information was being accumulated in both cases, given the very brief time that was available for actual stimulus processing on congruent flanker trials (i.e., 17 msec, plus 50 msec before the mask).

AUTHOR NOTE

This work was fully supported by Natural Sciences and Engineering Research Council of Canada (NSERC) Individual Discovery Grants to

both C.L.-S. and W.M.P. Address correspondence to either J. R. Schoenherr or C. Leth-Steensen, Department of Psychology, Carleton University, 1125 Colonel By Drive, Ottawa, ON, K1S 5B6 Canada (e-mail: psychophysics.lab@gmail.com or craig_leth_steensen@carleton.ca).

REFERENCES

- BARANSKI, J. V., & PETRUSIC, W. M. (1994). The calibration and resolution of confidence in perceptual judgments. *Perception & Psychophysics*, **55**, 412-428.
- BARANSKI, J. V., & PETRUSIC, W. M. (1995). The calibration of knowledge and perception. *Canadian Journal of Experimental Psychology*, **49**, 397-407.
- BARANSKI, J. V., & PETRUSIC, W. M. (1996). A doubt-scaling model of confidence calibration in human judgment [Abstract]. *International Journal of Psychology*, **31**, 186.
- BARANSKI, J. V., & PETRUSIC, W. M. (1998). Probing the locus of confidence judgments: Experiments on the time to determine confidence. *Journal of Experimental Psychology: Human Perception & Performance*, **24**, 929-945.
- BARANSKI, J. V., & PETRUSIC, W. M. (1999). Realism of confidence in sensory discrimination. *Perception & Psychophysics*, **61**, 1369-1383.
- BARANSKI, J. V., & PETRUSIC, W. M. (2001). Testing the architectures of the decision-confidence relation. *Canadian Journal of Experimental Psychology*, **55**, 195-206.
- BARANSKI, J. V., & PETRUSIC, W. M. (2003). Adaptive decision processes in perceptual comparisons: Effects of changes in the global difficulty context. *Journal of Experimental Psychology: Human Perception & Performance*, **29**, 658-674.
- BAR-TAL, Y., SARID, A., & KISHON-RABIN, L. (2001). A test of the overconfidence phenomenon using audio signals. *Journal of General Psychology*, **128**, 76-80.
- BILLET, M. T., & QIAN, Y. M. (2008). Are overconfident CEOs born or made? Evidence of self-attribution bias from frequent acquirers. *Management Science*, **54**, 1037-1051.
- BUSEY, T. A., TUNNICLIFF, J., LOFTUS, G. R., & LOFTUS, E. F. (2000). Accounts of the confidence-accuracy relation in recognition memory. *Psychonomic Bulletin & Review*, **7**, 26-48.
- CESARANI, D., LICHTENSTEIN, P., JOHANNESSON, M., & WALLACE, B. (2009). Heritability of overconfidence. *Journal of the European Economic Association*, **7**, 617-627.
- CHRIST, S., FALKENSTEIN, M., HEUER, H., & HOHNSEIN, J. (2000). Different error types and error processing in spatial stimulus-response-compatibility tasks: Behavioral and electrophysiological data. *Biological Psychology*, **51**, 129-150.
- CHRISTENSEN-SZALANSKI, J. J., & BUSHYHEAD, J. B. (1981). Physicians' use of probabilistic information in a real clinical setting. *Journal of Experimental Psychology: Human Perception & Performance*, **7**, 928-935.
- COLES, M. G. H., GRATTON, G., BASHORE, T. R., ERIKSEN, C. W., & DONCHIN, E. (1985). A psychophysiological investigation of the continuous flow model of human information processing. *Journal of Experimental Psychology: Human Perception & Performance*, **11**, 529-553.
- DHOLAKIA, U. M., & BAGOZZI, R. P. (2002). Mustering motivation to enact decisions: How decision process characteristics influence goal realization. *Journal of Behavioral Decision Making*, **15**, 167-188.
- EREV, I., WALLSTEN, T. S., & BUDESCU, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review*, **101**, 519-527.
- ERIKSEN, B. A., & ERIKSEN, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, **16**, 143-149.
- ERIKSEN, C. W., & HOFFMAN, J. E. (1972). Temporal and spatial characteristics of selective encoding from visual displays. *Perception & Psychophysics*, **12**, 201-204.
- ERIKSEN, C. W., & HOFFMAN, J. E. (1973). The extent of processing of noise elements during selective encoding from visual displays. *Perception & Psychophysics*, **14**, 155-160.
- FERRELL, W. R., & MCGOEY, P. J. (1980). A model of calibration for subjective probabilities. *Organizational Behavior & Human Performance*, **26**, 32-53.
- FISCHHOFF, B., SLOVIC, P., & LICHTENSTEIN, S. (1977). Knowing with

- certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception & Performance*, **3**, 552-564.
- GIGERENZER, G., HOFFRAGE, U., & KLEINBÖLTING, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, **98**, 506-528.
- GRIFFIN, D. W., & TVERSKY, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, **24**, 411-435.
- HEITZ, R. P., & ENGLE, R. W. (2007). Focusing the spotlight: Individual differences in visual attention control. *Journal of Experimental Psychology: General*, **136**, 217-240.
- JUSLIN, P., WINMAN, A., & OLSSON, H. (2000). Naive empiricism and dogmatism in confidence research: A critical examination of the hard-easy effect. *Psychological Review*, **107**, 384-396.
- KEREN, G. (1988). On the ability of monitoring non-veridical perceptions and uncertain knowledge: Some calibration studies. *Acta Psychologica*, **67**, 95-119.
- KEREN, G. (1991). Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologica*, **77**, 217-273.
- KLAPP, S. T., & HINKLEY, L. B. (2002). The negative compatibility effect: Unconscious inhibition influences reaction time and response selection. *Journal of Experimental Psychology: General*, **131**, 255-269.
- KORIAT, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review*, **100**, 609-639.
- KORIAT, A. (1995). Dissociating knowing and the feeling of knowing: Further evidence for the accessibility model. *Journal of Experimental Psychology: General*, **124**, 311-333.
- KVIDERA, S., & KOUTSTAAL, W. (2008). Confidence and decision type under matched stimulus conditions: Overconfidence in perceptual but not conceptual decisions. *Journal of Behavioral Decision Making*, **21**, 253-281.
- LABERGE, D. (1962). A recruitment theory of simple behavior. *Psychometrika*, **27**, 375-396.
- LAVIE, N. (1995). Perceptual load as a necessary condition for selective attention. *Journal of Experimental Psychology: Human Perception & Performance*, **21**, 451-468.
- LAVIE, N., & DE FOCKERT, J. W. (2003). Contrasting effects of sensory limits and capacity limits in visual selective attention. *Perception & Psychophysics*, **65**, 202-212.
- LICHTENSTEIN, S., & FISCHHOFF, B. (1977). Do those who know more also know more about how much they know? The calibration of probability judgments. *Organizational Behavior & Human Performance*, **20**, 159-183.
- LUCAS, J. A., & PETRUSIC, W. M. (2009). *Probing the locus of confidence in comparative judgments: Speed versus accuracy stress*. Manuscript under review.
- MERKLE, E. C., & VAN ZANDT, T. (2006). An application of the Poisson race model to confidence calibration. *Journal of Experimental Psychology: General*, **135**, 391-408.
- MURPHY, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology*, **12**, 595-600.
- OSKAMP, S. (1965). Overconfidence in case-study judgments. *Journal of Consulting Psychology*, **29**, 261-265.
- PAQUET, L., & CRAIG, G. L. (1997). Evidence for selective target processing with a low perceptual load flankers task. *Memory & Cognition*, **25**, 182-189.
- PAQUET, L., & LORTIE, C. (1990). Evidence for early selection: Precuing target location reduces interference from same-category distractors. *Perception & Psychophysics*, **48**, 382-388.
- PETRUSIC, W. M. (1992). Semantic congruity effects and theories of the comparison process. *Journal of Experimental Psychology: Human Perception & Performance*, **18**, 962-986.
- PETRUSIC, W. M., & BARANSKI, J. V. (1997). Context, feedback, and the calibration and resolution of confidence in perceptual judgments. *American Journal of Psychology*, **110**, 543-572.
- PETRUSIC, W. M., & BARANSKI, J. V. (2003). Judging confidence influences decision processing in comparative judgments. *Psychonomic Bulletin & Review*, **10**, 177-183.
- POLTROCK, S. E. (1989). A random walk model of numerical comparison. *Journal of Mathematical Psychology*, **33**, 131-162.
- RIDDERINKHOF, K. R. (2002a). Activation and suppression in conflict tasks: Empirical clarification through distributional analyses. In W. Prinz & B. Hommel (Eds.), *Attention and performance XIX: Common mechanisms in perception and action* (pp. 494-519). Oxford: Oxford University Press.
- RIDDERINKHOF, K. R. (2002b). Micro- and macro-adjustments of task set: Activation and suppression in conflict tasks. *Psychological Research*, **66**, 312-323.
- ROSLER, F., & FINGER, T. (1993). A psychophysiological analysis of response-channel activation and outcome states in Eriksen's noise-compatibility task. *Psychological Research*, **55**, 20-28.
- ROUDER, J. N., & KING, J. W. (2003). Flanker and negative flanker effects in letter identification. *Perception & Psychophysics*, **65**, 287-297.
- SANDERS, A. F., & LAMERS, J. M. (2002). The Eriksen flanker effect revisited. *Acta Psychologica*, **109**, 41-56.
- SMITH, J. D., & DUMONT, F. (2002). Confidence in psychodiagnostics: What makes us so sure. *Clinical Psychology & Psychotherapy*, **9**, 292-298.
- SOLL, J. B. (1996). Determinants of overconfidence and miscalibration: The roles of random error and ecological structure. *Organizational Behavior & Human Decision Processes*, **65**, 117-137.
- VAN LEEUWEN, C., & LACHMANN, T. (2004). Negative and positive congruence effects in letters and shapes. *Perception & Psychophysics*, **66**, 908-925.
- VAN ZANDT, T. (2000). ROC curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **26**, 582-600.
- VICKERS, D. (1979). *Decision processes in visual perception*. New York: Academic Press.
- WALLSTEN, T. S., & GONZALEZ-VALLEJO, C. (1994). Statement verification: A stochastic model of judgment and response. *Psychological Review*, **101**, 490-504.
- WALTER, S. D., COOK, D. J., GUYATT, G. H., SPANIER, A., JAESCHKE, R., TODD, T. R. J., & STREINER, D. L. (1998). Confidence in life-support decisions in the intensive care unit: A survey of healthcare workers. *Critical Care Medicine*, **26**, 44-49.

(Manuscript received June 5, 2009;
revision accepted for publication October 15, 2009.)