Journal of Experimental Psychology: Applied
2007, Vol. 13, No. 4, 182–196

1076-898X/07/$12.00 DOI: 10.1037/1076-898X.13.4.182

Fatigue, Sleep Loss, and Confidence in Judgment

Joseph V. Baranski
Defence Research and Development Canada

Sixty-four adults participated in a study examining the accuracy of metacognitive judgments during 28 hr of sleep deprivation (SD) and continuous cognitive work. Three tasks were studied (perceptual comparison, general knowledge, and mental addition), collectively spanning a range of cognitive abilities and levels of susceptibility to SD. Subjective and objective measures of sleepiness confirmed the expected patterns of increasing fatigue with SD. Participants displayed differing levels of metacognitive abilities across tasks, but traditional indices of the confidence–accuracy relation (i.e., calibration, resolution, over- and underconfidence), as well as the accuracy of pre- and posttask estimates of performance, remained stable over the SD period. The findings suggest that people can accurately assess their own cognitive performance when deprived of 1 night of sleep and that this ability need not be based on subjective estimates of sleepiness. The implications and limitations of the study are discussed and directions for future research are proposed.

Keywords: sleep deprivation, fatigue, confidence, calibration, metacognition

The antagonistic effects of sleep deprivation (SD) on human cognitive performance are well documented, and excellent reviews of various portions of this vast literature can be found in several sources (see Babkoff, Caspy, Mikulincer, & Sing, 1991b; Dinges & Kribbs, 1991; Harrison & Horne, 2000a; Horne, 1988a; Johnson, 1982; Kjellberg, 1977; Krueger, 1989; Pilcher & Huffcutt, 1996). The deleterious effects of even 1 night of sleep loss are most pronounced in simple vigilance, working memory, and psychomotor tasks (see, e.g., Angus & Heslegrave, 1985; Dinges & Kribbs, 1991; Wilkinson, 1965) and in subjective assessments of mood, alertness, and sleepiness (see, e.g., Babkoff, Caspy, & Mikulincer, 1991a; Gillberg, Kecklund, & Åkerstedt, 1994; Monk, 1987). However, more recent work has documented the effects of sleep loss in "higher level" cognitive tasks, such as those requiring creative problem solving, judgment, and decision making (see Harrison & Horne, 1998, 1999, 2000a; Horne, 1988b; Wimmer, Hoffmann, Bonato, & Moffitt, 1992; cf. Binks, Waters, & Hurry, 1999).

My sincere thanks to Tonya Hendriks, Andrea Hawton, and Heather Devine for their assistance with this study. I also thank Phillip Ackerman, Matthew Duncan, Keith Stewart, and Oshin Vartanian for helpful comments. Portions of this research were reported at the 50th Annual Meeting of the Human Factors and Ergonomics Society, San Francisco, California, October 2006.

Correspondence concerning this article should be addressed to Joseph V. Baranski, Collaborative Performance and Learning Section, Defence Research and Development Canada (Toronto), 1133 Sheppard Avenue West, Toronto, Ontario, Canada, M3M 3B9. E-mail: joe.baranski@drdc-rddc.gc.ca

One higher level capability that has received limited attention with respect to its susceptibility to SD is the "metacognitive" ability to self-monitor cognitive performance (i.e., the ability to judge how well or how poorly one is performing on a given cognitive task; e.g., Baranski, Pigeau, & Angus, 1994; Baranski et al., 2002; Baranski & Pigeau, 1997; Blagrove & Akehurst, 2000; Dorrian, Lamond, & Dawson, 2000). The issue is fundamentally important because fatigue due to sleep loss is so prevalent in contemporary societies (Coren, 1997; Dinges, 1995), and the consequences are potentially very serious, including loss of productivity (Krueger, 1989), occupational health and safety risks (Dawson & Fletcher, 2001), human error and accidents (Dinges, 1995; Horne & Reyner, 1995; Leger, 1994; Mitler et al., 1988; Webb, 1995), and reduced quality of life (Bonnet & Arand, 1995; Coren, 1997; Dement & Vaughan, 1999). Accordingly, it is important to understand the extent to which people can accurately evaluate their performance and cognitive abilities when sleep deprived as this can provide a basis for adaptive, compensatory behaviors. On the other hand, if this capability is vulnerable to the effects of SD, then it would provide a clear direction for public education, job- or task-specific training, and human factors intervention. Of course, an ability to self-monitor performance is necessary but not sufficient to preclude the antagonistic effects of SD; overt behavior is ultimately determined by many diverse goals and motivations, ill-advised as they might be at times (e.g., to continue driving late at night when very tired).

According to one perspective, it would be highly adaptive if the metacognitive ability to self-monitor performance was (at least slightly) more resistant to the effects of SD than primary cognitive task performance (Baranski et al., 1994); otherwise, there would be limited awareness of any decline in performance during SD. It is interesting that recent work has shown that tasks associated with the prefrontal cortex are particularly vulnerable to the effects of SD (Chee & Choo, 2004; Drummond et al., 1999, 2000; Durmer & Dinges, 2005; Harrison, Horne, & Rothwell, 2000; Horne, 1988a, 1993; Muzur, Pace-Schott, & Goldman-Rakic, 2002; Nilsson et al.,

2005; Thomas et al., 2000). Moreover, the metacognitive ability to self-monitor performance (i.e., metacognitive knowledge; see Fernandez-Duque, Baird, & Posner, 2000) is based on "executive functions" (Fernandez-Duque et al., 2000; Mazzoni & Nelson, 1998; Metcalfe & Shimamura, 1994; Shimamura, 2000), ostensibly involving the prefrontal cortex (see also Chua, Rand-Giovannetti, Schacter, Albert, & Sperling, 2004; Henson, Rugg, Shallice, & Dolan, 2000). From this perspective, then, the ability to self-monitor performance should be (at least) as vulnerable to SD effects as other prefrontal cognitive functions (Harrison & Horne, 2000a). Consistent with this view, Harrison and Horne (2000b) found that recognition memory for faces is not significantly impaired by more than 24 hr of SD, but temporal memory judgments for the same faces (i.e., whether they were presented early or late in the sequence), which involves the prefrontal cortex, were impaired by SD. More important for the purposes of the present study, participants in the SD conditions also displayed significantly higher confidence in incorrect judgments, relative to their non-SD counterparts (M = 2.26, SD = 0.74 vs. M = 1.55, SD = 0.65, on a scale of 1–5, where 1 = *just guessing* and 5 = *100% certain*). This suggests some antagonistic influence of SD on metacognitive functioning.

In comparison to this latter result, several studies have shown that sleep-deprived individuals are able to monitor their performance to a reasonable degree, although the findings to date have been based on global assessments made at the task level (Baranski & Pigeau, 1997; Dorrian et al., 2000) or on correlational indices that similarly provide a global view of the confidence–accuracy relationship (Baranski et al., 1994; Blagrove & Akehurst, 2000). In the present research, I provide an examination of judgments made at the task level, but my primary focus is on quantitative indices of self-monitoring based on trial-by-trial confidence ratings and "calibration" analyses.

Self-Monitoring Cognitive Performance: "Calibration" Research

The study of how accurately individuals can assess their own judgments and performance has been the focus of extensive research in the judgment and decision-making domain in the context of calibration studies (for reviews, see Harvey, 1997; Keren, 1991; Lichtenstein, Fischhoff, & Phillips, 1982; McClelland & Bolger, 1994; Suantak, Bolger, & Ferrell, 1996; Yates, 1990). In a typical calibration experiment, the participant makes repeated observations on some task and provides a confidence rating following each observation that reflects the degree of certainty in the judgment. The primary focus of calibration analyses is the proportion of correct judgments associated with each confidence level. Accurate self-monitoring of performance, and thus good calibration, is denoted by a monotone and rapidly increasing function relating confidence and judgment accuracy. Conversely, if the proportion of correct responses is relatively constant across the various confidence levels, then the participant cannot accurately self-monitor performance and, thus, is said to have poor calibration.

Trial-Level Confidence Judgments: Calibration, Resolution, and Over- and Underconfidence

Perhaps the most well-known index of trial-by-trial confidence judgments is over- and underconfidence (Lichtenstein & Fischhoff, 1977; Lichtenstein et al., 1982). In line with intuition, judgments are considered *overconfident* if the subjective proportion confidence exceeds the objective proportion correct on a given task; they are considered *underconfident* if the reverse is true (i.e., over-/underconfidence = $\bar{p} - \bar{e}$, where $\bar{p}$ denotes the mean proportion confidence and $\bar{e}$ denotes the mean proportion correct). Hence, a negative score denotes underconfidence and a positive score denotes overconfidence.

A second measure of the confidence–accuracy relation is *calibration*. Calibration refers to the correspondence between a probability assessment, expressed as the subjective probability of the occurrence of a particular event, and the empirical probability of the occurrence of that event. Hence, ideal calibration occurs when trials assigned a subjective probability of .5 are actually correct half of the time, .7 correct on judgments given a .7 subjective probability of being correct, and so on. Following Murphy (1973), the calibration score is denoted by a weighted squared deviation between the proportion correct associated with each confidence interval and the mean proportion confidence associated with each interval,

$$\frac{1}{n}\sum_{j=1}^{J} n_j(\bar{p}_j - \bar{e}_j)^2$$

where $\bar{p}_j$ and $\bar{e}_j$ are the mean proportion confidence and mean proportion correct in confidence interval j, respectively, n_j is the number of observations in confidence interval j, and n is the total number of observations. The calibration score ranges between 0.0 (optimal score) and 1.0 (the worst possible score), although in practice scores above 0.2 are rarely seen.

Finally, the Resolution index (Murphy, 1973) is denoted by a weighted squared deviation between the proportion correct associated with each confidence interval and the mean proportion correct,

$$\frac{1}{n}\sum_{j=1}^{J} n_j(\bar{e}_j - \bar{e})^2$$

In contrast to the calibration score, the resolution score provides an index of how well people use their confidence ratings to differentiate correct from incorrect responses (for a review, see Schneider, 1995). The resolution score likewise ranges between 0.0 and 1.0, although in this case 1.0 is optimal and 0.0 indicates no ability to distinguish correct from incorrect judgments. As with the calibration index, resolution scores above 0.2 are rarely seen. (For a more formal development of the indices associated with the accuracy of probability assessments, the reader is referred to the following sources: Baranski & Petrusic, 1994; Murphy, 1973; Yates, 1982, 1990; Yaniv, Yates, & Smith, 1991.) As mentioned, to date, the effects of sleep loss on trial-by-trial confidence ratings have been examined via correlational analyses (Baranski et al., 1994; Blagrove & Akehurst, 2000), but more formal, quantitative analyses of confidence calibration, resolution, and over-/underconfidence have not been examined.

Task-Level Estimates: Pre- and Posttask Judgments

Whereas confidence-based calibration analyses provide information concerning self-monitoring at the trial-by-trial level, a

more global level of analysis focuses on self-monitoring at the level of the task. Here, participants are asked to estimate, in terms of a proportion or percentage, their global level of performance. This is done before the task begins (i.e., a pretask estimate or predictive forecast) and after the task is completed (i.e., a posttask estimate or retrospective judgment). These estimates are then compared with actual task performance in terms of a global deviation, over-/underconfidence index (see also Baranski & Pigeau, 1997; Dorrian et al., 2000). The practical distinction between assessments made at the level of the trial (i.e., case-based, or probability assessments) and assessments made at the level of the task (i.e., category-based, or frequency assessments) is clear in terms of the nature of the judgments and the focus of analysis. For example, whereas confidence judgments capture the degree to which an individual can monitor trial-by-trial variation in performance, task-level assessments may reflect a broader consideration of personal and situational factors (for a review, see Ackerman & Wolman, 2007), such as consideration of one's current fatigue level or motivational state (pretask) or consideration of one's performance at time A relative to that of time B (posttask). In the present study, I examined both types of judgments as they reflect unique skills that have practical utility in the context of coping with fatigue due to sleep loss. (For a more complete discussion on the distinction between task and trial-level estimates, the interested reader is referred to the following sources: Brenner, Griffin, & Koehler, 2005; Brenner, Koehler, Liberman, & Tversky, 1996; Gigerenzer, Hoffrage, & Kleinbölting, 1991; Griffin & Buehler, 1999; Liberman, 2004.)

The Present Study

The objective of the present research was to examine the basis for the metacognitive ability to self-monitor cognitive performance during SD. According to one perspective, people have effective insight into their performance during SD (Baranski et al., 1994, 2002; Baranski & Pigeau, 1997). Although the absolute level of calibration may vary across tasks, it is assumed that calibration will remain unchanged as a function of SD. According to a second view, the ability to self-monitor performance is based in part on the subjective estimation of fatigue or sleepiness, and thus on an expectation of cognitive decline with increasing fatigue due to sleep loss (e.g., Blagrove & Akehurst, 2000; Dorrian et al., 2000). Although the role of subjective sleepiness in guiding overt behavior is clear (Monk, 1991), its role as the basis for self-monitoring is likewise pervasive and compelling: "In work situations with more or less passive supervision, such as driving tasks or control room tasks, the individual has no continuous feedback on the quality of performance. The subjective signals of sleepiness are the *only* [italics added] information on which the individual bases his decisions about when to discontinue work to avoid mistakes or accidents" (Gillberg, Kecklund, & Åkerstedt, 1994, p. 236). One way to assess this latter position is to examine tasks that vary in terms of their vulnerability to SD effects. Specifically, in Study 1, I examined three cognitive tasks, collectively spanning a range of cognitive abilities and levels of susceptibility to SD. The critical test of the latter hypothesis involves tasks that are less susceptible to SD effects: If performance assessments generally become more conservative with increasing fatigue levels, then tasks that show no or limited effects of sleep loss should display increased underconfidence as SD increases. Finally, to the extent that metacognition is regulated by executive functions under the control of the prefrontal cortex (Harrison & Horne, 2000a), such judgments should show a clear deterioration with increasing SD. Although, according to this latter position, it is unclear if this should manifest in increased over- or underconfidence, it is clear that poorer indices of calibration and resolution should be observed as SD increases.

Method

Participants

Sixty-four adults (67% male) volunteered for (approximately) 60 consecutive hr of experimentation.[1] Each participant was paid a fixed amount for his or her participation (approximately $400 CDN). Participants ranged in age from 17 to 33 years (M = 21.73 years, SD = 3.04); 67% reported being college or university educated. Participants were naive concerning the nature and aims of the study, but they were informed about the procedures to be employed, signed informed-consent forms for participation, and were debriefed on completion of the study.

Apparatus

Participants participated in groups of four but worked independently of each other and in their own rooms. Each experimental room was equipped with two computer workstations, a bed, and an intercom system that permitted communication with the experimenters. Electrophysiological recordings (i.e., EEG, EOG, EMG, and EKG) were collected with Embla Systems data recorders (details concerning the sleep physiology are reported in Baranski et al., 2007). Circadian rhythms were tracked by taking oral temperature every 2 hr using digital thermometers.

Design and Procedure

On arrival on Day 1, participants were familiarized with the lab, were briefed on the experiment, and completed a number of computer-based questionnaires, which were collected as part of a larger study. Following a lunch break, participants spent the remainder of Day 1 practicing the various individual and team tasks. During an afternoon break, participants were outfitted with the EEG recording equipment. The evening was spent relaxing, conversing, and watching movies. Participants slept in the lab from 2230 of Day 1 to 0630 of Day 2.

On awakening on Day 2, participants had breakfast, their electrodes were checked, and they prepared for the start of the experiment. The formal experiment ran continuously from 0800 on Day 2 to 1200 on Day 3. This 28-hr period consisted of 14 consecutive blocks, each of 2 hr in duration. Each 2-hr block comprised (approximately) 1 hr 40 min of sustained cognitive work and a 20-min break period. During breaks, participants ate, watched movies, used the washroom, had their electrodes checked, and interacted with each other and with the experimenters. During each 1-hr 40-min work period, participants performed three types of

[1] This article is based on data collected as part of a larger study on the effects of SD on individual and distributed team decision making. Results of the team decision-making data are reported in Baranski et al. (2007).

tasks: (a) a complex, distributed team decision-making task; (b) an individual-level cognitive task battery (i.e., six tasks involving numerical addition, perceptual comparison, general knowledge, air traffic control simulation, medical diagnosis, and spatial reasoning); and (c) an individual-level assessment battery that included a number of short tasks (e.g., simple reaction time, perceptual tracking, digit span, logical reasoning, etc.) and subjective questionnaires that probed mood, fatigue, and workload indices. Throughout the formal experiment, participants were unaware of the time of day (although they were asked for estimates in questionnaires delivered during the assessment battery) and decaffeinated foods and beverages were constantly available during breaks. The present study reports data from three tasks that were conducted during the individual-level cognitive testing sessions and permitted the collection of confidence judgments. Data collected during the team task focused on psychosocial aspects of distributed team decision making (i.e., social loafing and motivational gains) and are reported in a companion paper (Baranski et al., 2007).

The formal experiment ended at 1200 on Day 3. Participants were then given a 2-hr nap, followed by a short testing period that included a number of questionnaires. Participants were then debriefed, their electrodes were removed, and they were permitted to eat and shower. Participants were then released from the laboratory but were not permitted to drive themselves home (i.e., a ride was prearranged or a taxi was provided).

Cognitive Tasks

Three cognitive tasks were examined for calibration and self-monitoring analyses during the SD period: (a) a working memory task involving numerical addition, (b) a long-term memory task requiring answers to general knowledge questions, and (c) a visual, perceptual comparison task requiring relative judgments of line length. These tasks were chosen for a number of reasons. First, each has been employed in previous studies of confidence calibration (e.g., Baranski, Pigeau, Dinich, & Jacobs, 2004; Baranski & Petrusic, 1998, 1999, 2001). Second, they collectively span a wide range of cognitive abilities (i.e., short- and long-term memory processes and direct perceptual processes). Finally, and perhaps most important, they vary considerably in terms of their susceptibility to SD effects (see below).

Long-term memory task: General knowledge. The general knowledge task required answers to two-alternative trivia questions covering various topics (e.g., history, geography, literature, sports, etc.). Variations of this task have been used extensively in calibration studies in the judgment and decision-making literature, but the task has not been previously employed in a sleep loss study. The questions were selected from a large pool of items that were pretested (on 10 non–sleep-deprived adults) to exclude tricky or "misleading" questions (see Juslin, 1993, 1994). The questions were then randomly assigned to three test banks, each with approximately 100 questions. The order of presentation of the test banks was counterbalanced across participants and sessions. Each experimental trial comprised a question (e.g., "Which river is longer?") and two alternatives (e.g., "Nile" "Amazon"). The alternatives appeared side by side on the computer monitor and were placed directly below the question. Participants responded by depressing either the left or the right response key on the mouse to select the alternative they considered to be the correct response. Following each decision response, the screen was cleared and participants were prompted to select one of six confidence ratings, from 50% (*guess*) to 100% (*certain*) in steps of 10%.

Thirty-two participants performed the general knowledge task for exactly 24 min at the following 8-hr intervals during the SD phase of the experiment: 6, 14, and 22 hr of sleep loss. On average, each participant completed approximately 60 trials per session (M = 60.47, SD = 10.10) in each of the three sessions. Participants were required to respond as accurately as possible without taking too much time to respond. Feedback on the accuracy of the judgments or the validity of the confidence ratings was not provided at any point during the study.

Visual perceptual comparison task. The visual comparison task required the relative judgment of line length. This task has been used previously in calibration (e.g., Baranski & Petrusic, 1999; Juslin & Olsson, 1997) and sleep deprivation (e.g., Baranski et al., 2002; Baranski & Pigeau, 1997) studies. Each trial began with the presentation of an instruction ("LONGER" or "SHORTER"), which was displayed near the top of the computer monitor. One second later, a visual display appeared that consisted of two horizontal lines divided by one short vertical line. The display remained on the screen until the participant responded. The participant's task was to determine which of the two horizontal lines was longer or shorter, depending on the instruction. Participants responded by depressing either the left or the right button on the mouse to indicate that the left or right line was the longer or the shorter. Three levels of judgment difficulty were randomly presented to the participants; the difficulty was defined a priori on the basis of the ratio of the longer to the shorter line: 1.01, 1.03, and 1.05. All lines appeared black on a white background. As in the general knowledge task, each trial was followed by a confidence rating from 50% (*guess*) to 100% (*certain*) in steps of 10%.

Thirty-two participants performed the visual comparison task for exactly 24 min at the following 8-hr intervals during the SD phase of the experiment: 8, 16, and 24 hr of sleep loss. On average, each participant completed approximately 150 trials per session (M = 148.29, SD = 57.42) in each of the three sessions. Participants were required to respond as accurately as possible without taking too much time to respond. Feedback on the accuracy of the judgments or the validity of the confidence ratings was not provided at any point during the study.

Short-term memory task: Mental addition. This task is based on a similar task employed by Wilkinson (1969). The participant is required to mentally add a sequence of eight numbers presented on the computer monitor at a rate of one number every 1.25 s (see Baranski et al., 2002; Baranski & Pigeau, 1997). The numbers ranged from 1 to 16, and the sequence was terminated by the presentation of a visual prompt (=) at which time participants typed in their response and then pressed the "Enter" key. The number "10" was not used because of the disproportionate ease with which decades are added. Following each trial, the participants were prompted to key in a confidence rating, from 0 to 100, to reflect the degree of certainty in their response in terms of a subjective probability that their response was correct. Thus, a rating of 0 was to denote absolute certainty that their answer was wrong, and a rating of 100 was to denote absolute certainty that their answer was correct. Confidence ratings between 0 and 100 corresponded to increasing certainty in the correctness of the response in terms of a likelihood or percentage. Participants de-

pressed the "Enter" key to record their confidence ratings and to proceed to the next trial.

Note that the mental addition task required a "full-range" confidence scale, from 0% to 100%, whereas the general knowledge task and the perceptual comparison task required a "half-range" confidence scale from 50% to 100% (see Harvey, 1997). The difference lies in the fact that the latter tasks involve two-alternative forced-choice responses; thus, it is expected that the lowest level of response accuracy will correspond to chance responding (i.e., "guessing") and thus 50% accuracy. In the case of the addition task, on the other hand, participants can be certain of an error (e.g., they lose track of the counts, miss a number, or simply cannot perform the mental addition), and thus can (and do) report subjective certainty of an error response (i.e., 0% confidence in a correct response). The difference between the two types of tasks and subjective probability scales was explained in great detail to the participants, both in the case of trial-by-trial confidence ratings and also in terms of pre- and posttask estimates of performance.

Twenty participants performed the mental addition task for exactly 24 min at the following 8-hr intervals during the SD phase of the experiment: 4, 12, 20, and 28 hr of sleep loss. On average, each participant completed approximately 50 trials per session ($M = 50.27$, $SD = 6.32$) in each of the four sessions. Participants were required to respond as accurately as possible without taking too much time to respond. Feedback on the accuracy of the judgments or the validity of the confidence ratings was not provided at any point during the study.

Pre- and posttask estimates. To assess at the task level the extent to which participants were able to accurately assess their own cognitive abilities during SD, I asked them to estimate the percentage of responses that they thought they would answer correctly (i.e., pretask estimate) before each of the three aforementioned tasks. In addition, each task was followed by a similar question that asked participants to estimate the percentage of responses that they answered correctly (i.e., posttask estimate). Participants' self-monitoring ability was assessed by comparing the task-level estimates of performance with actual performance accuracy. When the estimates exceeded or fell below actual performance, I concluded that participants were overconfident or underconfident in their assessments, respectively. When assessments closely match performance, I concluded that participants were "well-calibrated" in their task-level estimates.

Susceptibility of the Tasks to SD

As discussed in the introduction, cognitive tasks involving working memory or monotonous tasks are susceptible to the effects of SD. For the current version of the mental addition task, which involves a substantial working memory component, previous studies have found moderate impairments in task performance over 24 hr of SD (e.g., 15%–30% decrease in accuracy; Baranski & Pigeau, 1997; Baranski et al., 1994, 2002). The perceptual comparison task, in contrast, involves limited working memory but requires vigilance and thus may be vulnerable to lapses in attention or motivation. Previous studies employing this task have found small (5%–7%) but significant declines in accuracy over 24 hr of SD (Baranski & Pigeau, 1997; Baranski, Cian, Esquivié, Pigeau, & Raphel, 1998). Finally, the general knowledge task, as mentioned, has not been used in previous SD studies. However, the task is limited in terms of vigilance requirements and working memory processes; thus, performance is not expected to show significant deterioration with sleep loss. Nevertheless, for each task, the critical questions concern the effects of sleep loss on confidence, calibration, resolution, over- and underconfidence, and task-level self-monitoring.

Results

The results are presented in four sections. The first reports oral temperatures and subjective sleepiness scores with a view toward establishing that participants showed the expected patterns during the SD period. Sections 2–4 provide analyses of task performance (i.e., response accuracy) and the confidence, calibration, resolution, over- and underconfidence, and global task-level indices of self-monitoring for the general knowledge, line comparison, and mental addition tasks, respectively.

The results and analyses treat each task separately because the counterbalancing schedule of the individual cognitive tasks necessitated that the tasks were conducted at different times of the day. Moreover, given the task rotation schedule, it was necessary to counterbalance tasks across participants, with the result that some participants performed only one of the three tasks, some performed two, but none performed all of the tasks. For each task, I performed a repeated measures analysis of variance (ANOVA) with session as a within-participants factor. Significance values were based on the Huynh–Feldt adjusted degrees of freedom; these adjusted degrees of freedom are reported in Tables 1–3. An alpha level of .05 was adopted throughout. Statistical power analyses were based on Cohen (1988) and were calculated using G*Power algorithms (Buchner, Erdfelder, & Faul, 1997) assuming a medium effect size for repeated measures designs (effect size = 0.15).

The analyses for the sleepiness and temperature measures were based on the data of all 64 participants. However, the cognitive task data of some participants were excluded because of poor performance (i.e., response accuracy). Specifically, I excluded 1 participant on the addition task (resulting $n = 19$), 2 participants on the knowledge task (resulting $n = 30$), and 3 participants on the

Table 1
ANOVA Statistics for the General Knowledge Task

Variable	Source	*df*	*F*	Partial η^2
Accuracy	Session	1.94	0.23	—
	Error	56.28	(.00)	
Confidence	Session	1.76	2.36	.08
	Error	50.92	(15.15)	
Under-/overconfidence	Session	1.92	0.98	—
	Error	55.76	(52.30)	
Calibration	Session	1.88	2.22	.07
	Error	54.55	(.00)	
Resolution	Session	1.73	0.20	—
	Error	50.06	(.00)	
Pretask accuracy estimate	Session	1.90	1.18	.01
	Error	54.95	(64.71)	
Posttask accuracy estimate	Session	2.00	0.56	—
	Error	58.00	(50.05)	

Note. Values in parentheses represent mean square errors. Dashes indicate that effect sizes are not provided for F values < 1.0.

Table 2
ANOVA Statistics for the Perceptual Comparison Task

Variable	Source	*df*	*F*	Partial η^2
Accuracy	Session	1.68	3.79*	.12
	Error	49.60	(39.28)	
Confidence	Session	1.75	7.44**	.21
	Error	48.89	(47.23)	
Under-/overconfidence	Session	1.75	0.86	—
	Error	49.10	(60.85)	
Calibration	Session	2.00	0.16	—
	Error	56.00	(.00)	
Resolution	Session	1.24	0.96	—
	Error	34.74	(.00)	
Pretask accuracy estimate	Session	1.93	2.13	.07
	Error	54.00	(92.14)	
Posttask accuracy estimate	Session	1.66	0.24	—
	Error	46.54	(81.84)	

Note. Values in parentheses represent mean square errors. Dashes indicate that effect sizes are not provided for F values < 1.0.
* $p < .05$. ** $p < .01$.

perceptual task (resulting $n = 29$), after an examination of the individual participant data revealed patterns of fast guessing (i.e., random responding) during portions of the sessions.

All figures reporting error bars show the mean and 95% confidence intervals (see Masson & Loftus, 2003). Figures reporting calibration functions are based on the overall data set and show the frequency with which the various confidence levels were employed.

Sleepiness and Temperature Measures

As mentioned, oral temperatures and Stanford Sleepiness Scores (SSS; Hoddes, Dement, & Zarcone, 1972) were obtained every 2 hr during the formal experiment. Oral temperature is an established circadian correlate of fatigue and performance (see Froberg, 1977; Monk, 1991; Monk, Leng, Folkard, & Weitzman, 1983), and the SSS is a validated sleepiness index employed in numerous sleep loss and performance studies (Babkoff et al., 1991a; Glenville & Broughton, 1978; Hoddes, Zarcone, Smythe, Phillips, & Dement, 1973; Moses, Lubin, Naitoh, & Johnson, 1978). For example, Glenville and Broughton (1978) found significant correlations ($r = .50 - .69$) between SSS and performance on tests of vigilance, choice reaction, and simple reaction time. The scale ranges from 1 (*feeling active and vital; alert; wide awake*) to 7 (*almost in reverie; sleep onset soon; losing struggle to remain awake*).

Table 3
ANOVA Statistics for the Mental Addition Task

Variable	Source	*df*	*F*	Partial η^2
Accuracy	Session	3.00	8.29**	.32
	Error	54.00	(98.45)	
Confidence	Session	2.14	9.04**	.33
	Error	38.49	(221.80)	
Under-/overconfidence	Session	2.00	2.01	.13
	Error	36.00	(88.11)	
Calibration	Session	2.53	1.79	.09
	Error	45.59	(.00)	
Resolution	Session	2.11	0.31	—
	Error	37.97	(.00)	
Pretask accuracy estimate	Session	2.54	1.27	.07
	Error	45.72	(239.00)	
Posttask accuracy estimate	Session	2.64	1.17	.06
	Error	47.50	(228.33)	

Note. Values in parentheses represent mean square errors. Dashes indicate that effect sizes are not provided for F values < 1.0.
* $p < .05$. ** $p < .01$.

Figure 1 provides plots of oral temperature and SSS throughout the formal experiment (i.e., during Day 2 and Day 3 of the study. Recall that Day 1 was for familiarization and task practice). The figure also notes the times at which the various cognitive tests were conducted (e.g., the addition task was conducted at 1030 and 1830 on Day 2 and at 0230 and 1030 of Day 3). These data display the typical circadian patterns that have been well established in numerous sleep loss and performance studies (e.g., see Monk, 1987), and they are also typical of the task subsamples in the present study: Temperature decreases, $F(1, 63) = 15.66$, $p < .01$, $MSE = 0.09$, partial $\eta^2 = .20$, and sleepiness increases, $F(1, 63) = 61.85$, $p < .01$, $MSE = 1.01$, partial $\eta^2 = .50$, with increasing SD. Most important for the purposes of the present study, oral temperatures were lowest and sleepiness scores were highest for the third and final testing session of the perceptual and general knowledge tasks; for the addition task, oral temperatures were low and sleepiness scores were high for Sessions 3 and 4. Taken together, the results in Figure 1 confirm that participants were quite tired by the time of the final testing sessions.

General Knowledge Task

Figure 2 shows the calibration curve for the general knowledge task. The calibration curve plots the percentage of correct responses associated with each level of confidence reported by the participants (the numbers beside each point denote the percentage of times that each confidence level was used by participants). Accordingly, perfect calibration would be characterized by points falling along the main diagonal; under- and overconfidence would be characterized by points above and below the diagonal, respectively. Note that the range of the calibration curve is between 50% and 100% for this two-alternative, forced choice general knowledge task. That is, the lowest confidence rating is 50% or "guessing," which reflects that this task was not expected to produce a response accuracy level below 50% (unless, of course, the questions were tricky or misleading; see Baranski & Petrusic, 1995; Juslin, 1994).

The results in Figure 2 show that participants were generally overconfident on the general knowledge task (i.e., a considerable portion of the data points fall below the identity line). For example, participants reported certainty (i.e., 100% confidence) on 23.6% of the trials, but they were in fact only 66% correct on those trials. It is important to note that, although there have been occasional reports of good calibration in the general knowledge domain (Gigerenzer et al., 1991; Juslin, 1993, 1994), the finding of overconfidence is typical when judgments are difficult; thus, the percentage of correct responses is low (Baranski & Petrusic, 1995, 2001; Blais, Thompson, & Baranski, 2005; Brenner et al., 1996; Ferrell & McGoey, 1980; Griffin & Buehler, 1999; Griffin & Tversky, 1992; Harvey, 1997; McClelland & Bolger, 1996; Suantak, Bolger, & Ferrell, 1996). Indeed, the percentage of correct

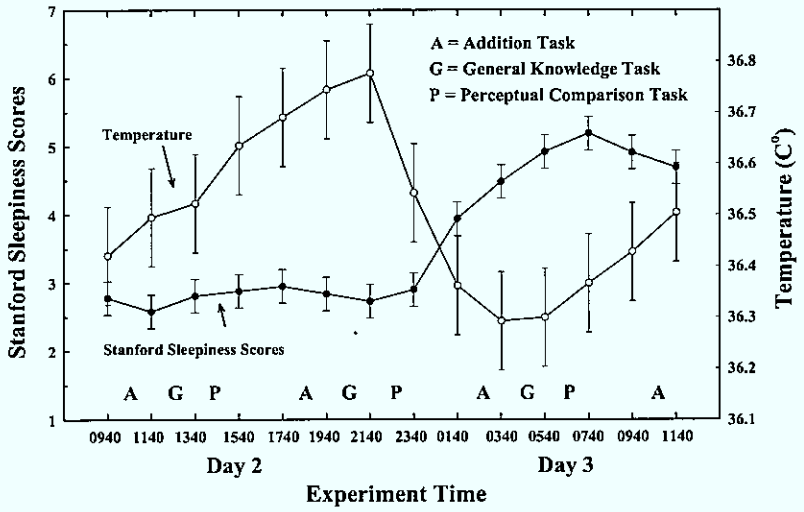

Figure 1. Mean Stanford Sleepiness Scores (filled circles) and oral temperatures (open circles) plotted as a function of time of day during the experimental portion of the study. Also shown are the various times at which the three cognitive tasks were performed. The error bars denote 95% confidence intervals across all 64 participants. The figure is adapted from Baranski et al. (2007, Figure 1).

responses in the present study was 63%; thus, the finding of overconfidence is not unexpected. Finally, although not shown, the calibration curve remains relatively constant across sessions; for clarity, only the overall plot is provided.

The main findings are shown in Figures 3 and 4 and a summary of the various ANOVAs is provided in Table 1. Figure 3A provides a plot of the mean percentage of correct responses and mean confidence ratings across the three sessions of the general knowledge task (i.e., at 6, 14, and 22 hr of sleep loss). The plot shows that mean judgment accuracy and confidence do not significantly change as a function of hours of sleep loss (power = 0.97 and 1.00, respectively). The difference between confidence and response accuracy provides the index of over-/underconfidence and is reported in Figure 3B. As the figure shows, participants were approximately 10% overconfident, $t(29) = 4.55$, $p < .01$, $d = 0.83$, with these relatively difficult knowledge questions. However, as is

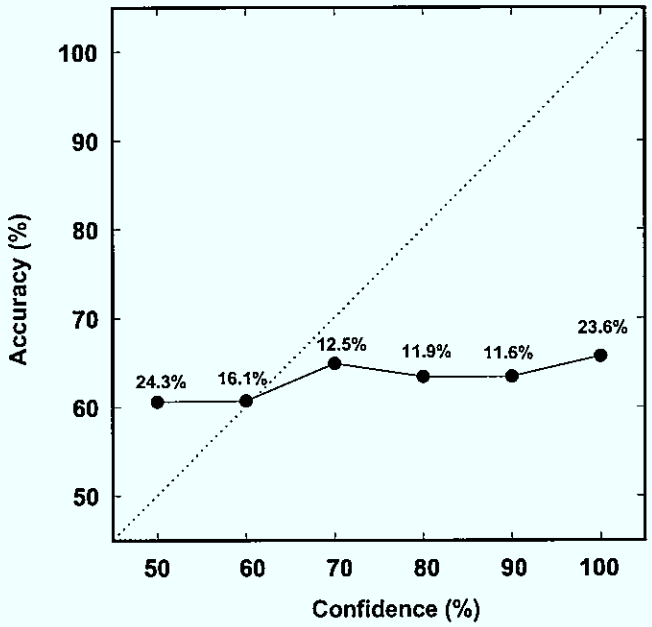

Figure 2. Calibration curve for the general knowledge task. The main diagonal denotes perfect calibration. The percentage of time each confidence category was used is provided beside each data point. The figure is based on 5,279 observations.

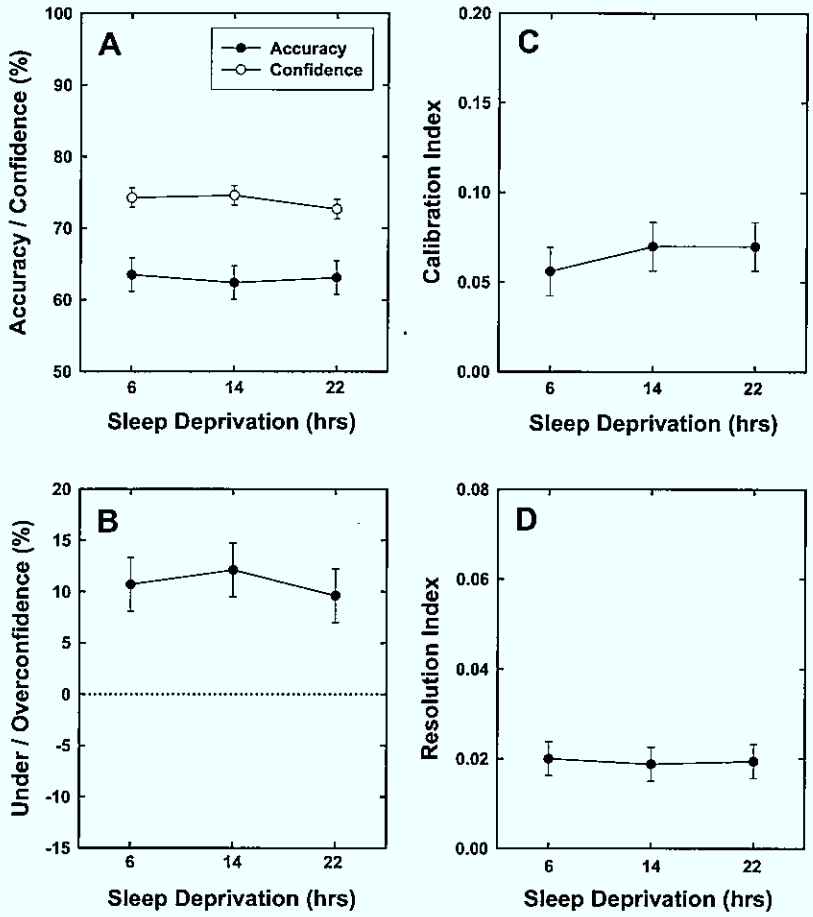

Figure 3. Mean response confidence and response accuracy (A), under-/overconfidence (B), calibration (C), and resolution (D) plotted as a function of hours of sleep loss for the general knowledge task. Error bars denote 95% confidence intervals.

also evident in Figure 3B, the degree of overconfidence did not significantly change with increasing sleep loss (power = 1.00). Finally, Figures 3C and 3D provide plots of the calibration and resolution indices across the three sessions. As was the case with the other measures, calibration and resolution did not significantly change as a function of sleep loss (power = 0.99 and 0.95, respectively).

Figure 4 provides a view of the ability to judge performance at the task level. Specifically, pretask (open circles) and posttask (filled triangles) estimates of performance are provided with the actual percentage of correct responses (filled circles). Note that although plotted at the same time, the pre- and posttask estimates were actually obtained just prior to commencing the task and immediately following the completion of the task, respectively (similarly for Figures 7 and 10). The plot shows that participants were, once again, approximately 10% overconfident in their pretask estimates, $t(29) = 4.32$, $p < .01$, $d = 0.79$, and their posttask estimates, $t(29) = 3.90$, $p < .01$, $d = 0.71$. However, these differences did not change significantly as a function of SD (i.e., participants became neither more overconfident nor more underconfident as their fatigue levels increased; power = 0.98 and 1.00, respectively).

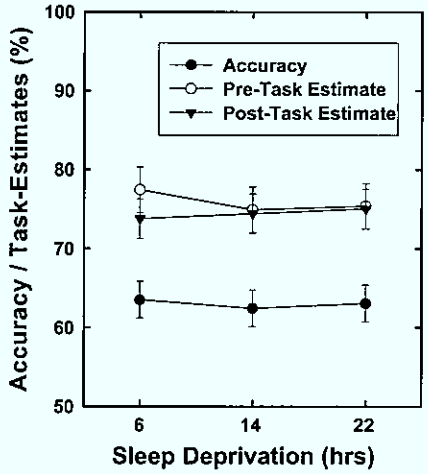

Figure 4. Mean response accuracy (filled circles), pretask estimates (open circles), and posttask estimates (filled triangles) plotted as a function of hours of sleep loss for the general knowledge task. Error bars denote 95% confidence intervals.

It is interesting that the pre- and posttask estimates were significantly correlated across participants (*r*s = .47, .66, and .75, for Sessions 1–3, respectively, all *p*s < .05, *df* = 28), suggesting that the participants who made higher pretask estimates tended to make higher posttask estimates. However, participants who made higher estimates were not more likely to be more accurate; that is, the pre- and posttask estimates were not significantly correlated with response accuracy. These results are consistent with studies examining individual differences in confidence and accuracy in two-alternative forced-choice tasks (for a review, see Blais et al., 2005), but they are not consistent with studies examining the validity of self-estimates of intellectual abilities (for a review and study, see Ackerman & Wolman, 2007). One fundamental difference between these lines of research is that in the former tasks participants are typically asked to estimate their own performance directly, whereas in the latter studies participants are typically asked to estimate their performance relative to other participants (e.g., in terms of a percentile ranking).

In summary, the data in Figure 1 confirmed that participants' fatigue and sleepiness levels increased with SD. In terms of performance, accuracy and confidence did not change significantly with increasing SD. Although participants were overconfident in their assessments on a difficult general knowledge task, the indices of self-monitoring were not significantly affected by SD. In addition, the absence of a trend toward underconfidence in the trial confidence and pre- and posttask estimates suggests that participants need not base their metacognitive judgments on assessments of subjective fatigue (e.g., Blagrove & Akehurst, 2000; Dorrian et al., 2000; Gillberg et al., 1994).

Perceptual Comparison Task

Figure 5 shows the calibration curve for the perceptual comparison task. As was the case in the general knowledge task, participants showed overconfidence over most of the confidence levels in this relatively difficult two-alternative, forced-choice task.

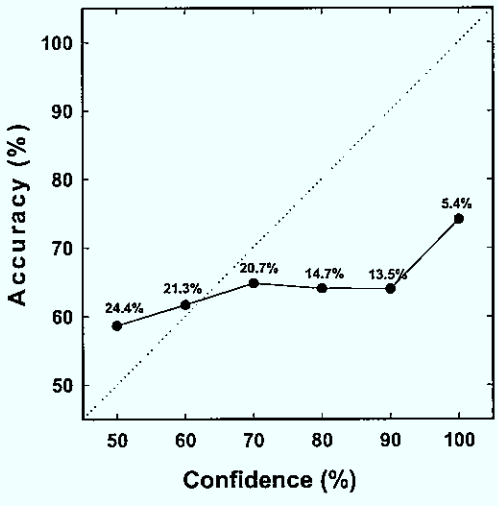

Figure 5. Calibration curve for the perceptual comparison task. The main diagonal denotes perfect calibration. The percentage of time each confidence category was used is provided beside each data point. The figure is based on 12,814 observations.

Again, although there have been reports of good calibration, or even underconfidence (Baranski & Petrusic, 1994; Björkman, Juslin, & Winman, 1993; Juslin & Olsson, 1997) in the domain of perceptual judgments, the finding of overconfidence is typically observed when judgments are difficult and, thus, the percentage of correct responses is low (Baranski & Petrusic, 1994, 1999; Petrusic & Baranski, 1997). Indeed, the percentage of correct responses in the present study was 63% and, thus, the finding of overconfidence is not unexpected.

The main findings are shown in Figures 6 and 7 and a summary of the various ANOVAs is provided in Table 2. Figure 6A provides a plot of the mean percentage of correct responses and mean confidence ratings across the three sessions of the perceptual comparison task (i.e., at 8, 16, and 24 hr of sleep loss). The plot shows that judgment accuracy demonstrates a small (4%) but significant decline with increasing SD. Mirroring response accuracy, mean response confidence likewise shows a small but significant decline with increasing sleep loss. Figure 6B plots the difference between confidence and response accuracy. Overall, participants were approximately 5% overconfident on the perceptual task, but this difference was not significant, $t(28) = 1.83$, *ns*, $d = 0.34$. More important, the level of overconfidence did not significantly change as a function of SD (power = 1.0). Finally, Figures 6C and 6D show that calibration and resolution likewise did not change significantly as a function of SD (power = 0.99 and 0.96, respectively).

Figure 7 provides a view of the ability to judge performance at the task level. Specifically, pretask (open circles) and posttask (filled triangles) estimates of performance are provided with the actual percentage of correct responses (filled circles). As in the general knowledge task, the pre- and posttask estimates were again significantly correlated across sessions (*r*s = .64, .42, and .68, for Sessions 1–3, respectively, all *p*s < .05, *df* = 27), but they were not significantly correlated with accuracy across participants.

As is evident in Figure 7, participants were generally well calibrated at the task level; overall, the differences between the estimates and actual accuracy were not statistically different from zero for the pretask estimates, $t(28) = 1.58$, *ns*, $d = 0.29$, and the posttask estimates, $t(28) = 0.01$, *ns*, $d = 0.00$. More important, this level of self-monitoring again did not change significantly as a function of SD (power = 0.92 and 0.99 for the pre- and posttask estimates, respectively).

In summary, the perceptual comparison showed a small but significant decline in response accuracy with increasing sleep loss. However, confidence judgments and task-level estimates mirrored this decline, resulting in no significant change in over-/underconfidence with increasing SD. These findings demonstrate that the ability to monitor performance during SD is sensitive to even relatively small changes in task performance.

Mental Addition Task

Figure 8 shows the calibration curve for the mental addition task. In contrast to the general knowledge and perceptual comparison tasks, the mental addition task requires a "full-range" calibration curve analysis (see Harvey, 1997). Specifically, in the present context and as discussed earlier, participants can indeed be certain of an error (i.e., 0% confidence), and thus the use of a full-range confidence scale is essential. Note that in contrast to the general

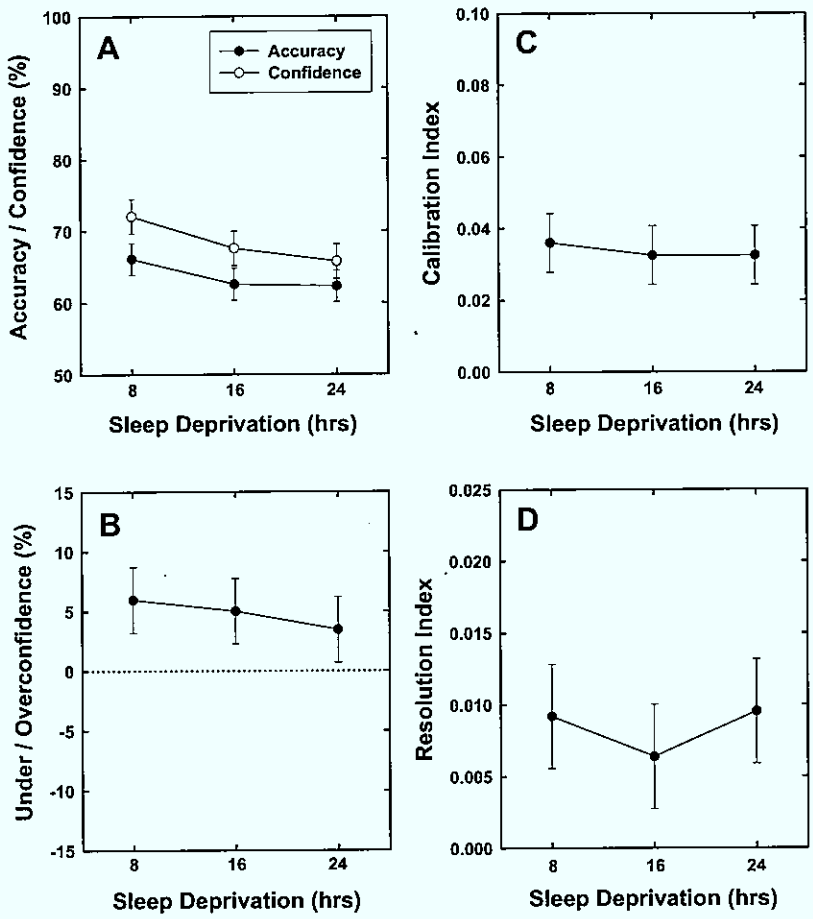

Figure 6. Mean response confidence and response accuracy (A), under-/overconfidence (B), calibration (C), and resolution (D) plotted as a function of hours of sleep loss for the perceptual comparison task. Error bars denote 95% confidence intervals.

knowledge and perceptual comparison tasks, participants demonstrated remarkable trial-by-trial calibration in the mental addition task (see also Baranski et al., 1994, 2004). Indeed, across sessions and responses, participants were nearly perfectly calibrated in their confidence judgments.

The main findings are shown in Figures 9 and 10, and a summary of the various ANOVAs is provided in Table 3. Figure 9A provides a plot of the mean percentage of correct responses and mean confidence ratings across the four sessions of the addition task (i.e., at 4, 12, 20, and 28 hr of sleep loss). The plot shows that judgment accuracy demonstrated a relatively large (≥ 15%) and significant decline with increasing SD, mirrored by an appropriately large and significant decline in response confidence. Figure 9B plots the difference between confidence and response accuracy. Overall, participants were neither over- nor underconfident in their confidence judgments, $t(18) = -0.64$, *ns*, $d = 0.00$. Although Figure 9B suggests a tendency toward underconfidence with increasing SD, the effect of session was not statistically significant (power = 0.98). Finally, Figures 9C and 9D show that calibration and resolution once again did not change significantly as a function of sleep loss (power = 0.98 and 0.98, respectively).

Figure 10 provides a view of the ability to judge performance at the task level. As with the other tasks, pre- and posttask estimates

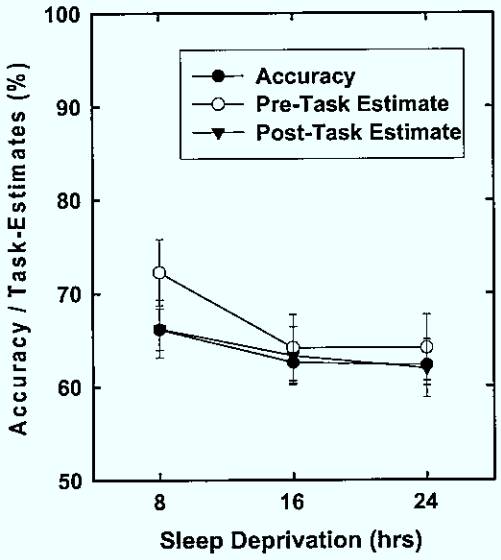

Figure 7. Mean response accuracy (filled circles), pretask estimates (open circles), and posttask estimates (filled triangles) plotted as a function of hours of sleep loss for the perceptual comparison task. Error bars denote 95% confidence intervals.

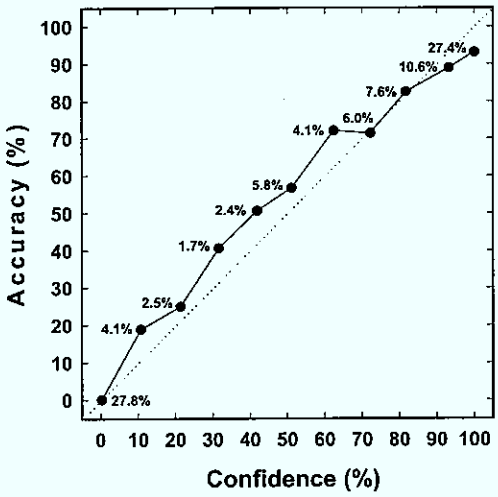

Figure 8. Calibration curve for the mental addition task. The main diagonal denotes perfect calibration. The percentage of time each confidence category was used is provided beside each data point. The figure is based on 3,750 observations.

were significantly correlated (rs = .97, .86, .58, and .76 for Sessions 1–4, respectively, all $ps < .05$, $df = 17$). Unlike with the other tasks, however, significant ($p < .05$, $df = 17$) correlations were observed between the pre- and posttask estimates and accuracy across participants for Session 1 (rs = .68 and .73, respectively), Session 2 (rs = .53 and .75, respectively), Session 3 (rs = .29, *ns*, and .47, respectively) and Session 4 (rs = .46 and .50, respectively).

Overall, the difference between the pretask estimates and accuracy was not significantly different from zero, $t(18) = -1.32$, *ns*, $d = -0.30$, but the difference between the posttask estimates and accuracy was significantly different from zero in the direction of underconfidence, $t(18) = -2.34$, $p < .05$, $d = -0.54$. More important, in both cases, participants became neither more overconfident nor more underconfident with increasing SD; that is, the main effect of session was not significant (power = 0.98 and 1.0, respectively).

In summary, participants were extremely well calibrated on the mental addition task. Although response accuracy displayed a relatively large and significant decline with increasing SD, response confidence and task-level estimates mirrored this decline, resulting in no significant effects of SD on the various self-

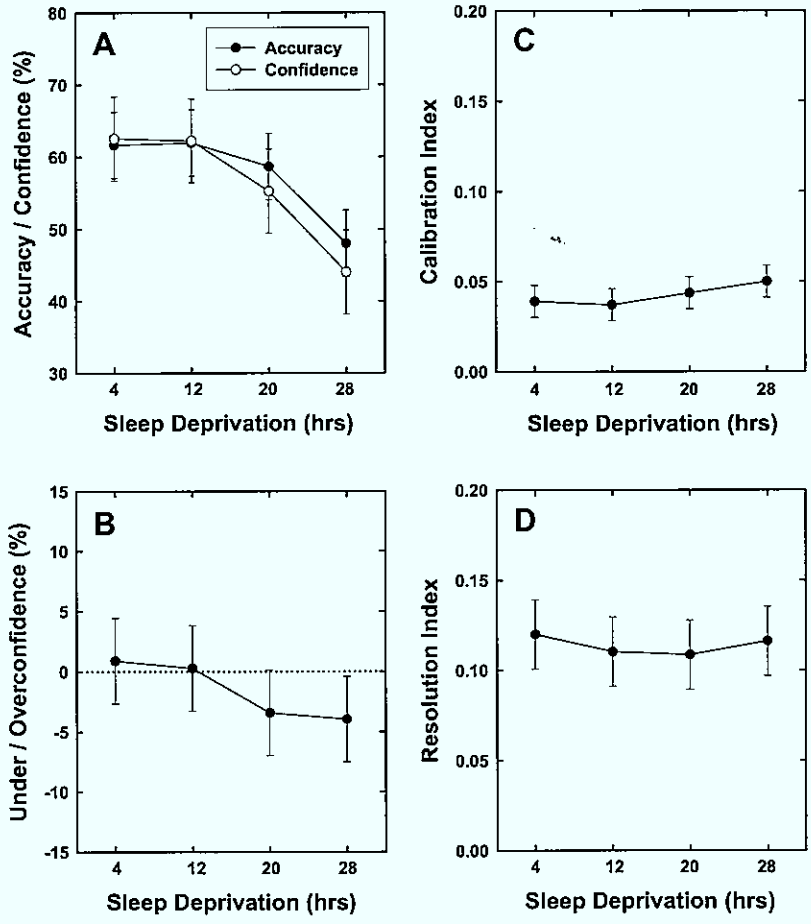

Figure 9. Mean response confidence and response accuracy (A), under-/overconfidence (B), calibration (C), and resolution (D) plotted as a function of hours of sleep loss for the mental addition task. Error bars denote 95% confidence intervals.

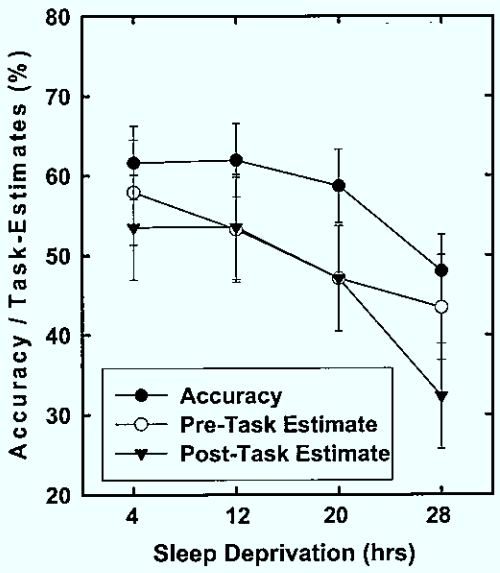

Figure 10. Mean response accuracy (filled circles), pretask estimates (open circles), and posttask estimates (filled triangles) plotted as a function of hours of sleep loss for the mental addition task. Error bars denote 95% confidence intervals.

monitoring indices. These findings indicate that participants were able to self-monitor relatively large changes in cognitive task performance during SD.

Discussion

The findings of the present study replicate previous reports of effective task-level, metacognitive ability during 24 hr of SD (Baranski et al., 2002; Baranski & Pigeau, 1997; Dorrian et al., 2000), and they extend these findings to include trial-by-trial confidence calibration analyses. In three tasks, covering a wide range of cognitive abilities and levels of susceptibility to SD, the ability to assess accurately one's own performance did not significantly deteriorate over the SD period. Although the findings taken together are generally encouraging with respect to the robustness of metacognitive abilities, several of the results from the specific tasks are suggestive in their own respect. I now address these issues in turn.

Subjective Sleepiness and Metacognition During Sleep Deprivation

One objective of the present study was to distinguish the self-monitoring of task performance from the self-monitoring of fatigue or sleepiness. In this case, the findings from the general knowledge task were most relevant because task accuracy was not significantly affected by SD. Specifically, if confidence judgments and task-level assessments are generally influenced by fatigue or sleepiness levels, then increasing levels of SD should have led to an increase in underconfidence. That this was not the case suggests that subjective estimates of fatigue and sleepiness need not provide the basis for the self-monitoring of performance (Blagrove & Akehurst, 2000; Dorrian et al., 2000; Gillberg et al., 1994); evidently, participants can monitor their task performance, independent of subjective sleepiness levels.

Of course, this is not to say that subjective assessments of fatigue, sleepiness, or motivation are never used. Indeed, it is certain that such assessments provide critical cues for guiding behavior during SD (see Monk, 1991), especially when performing an unfamiliar task or in a context in which there is no or limited explicit feedback from the environment. When performing a more well-practiced task, however, it is possible to monitor aspects of performance directly or, more specifically, to monitor systematic changes in aspects of performance. For example, in the absence of explicit feedback from the task or environment, one aspect of performance that can be continuously monitored is subjective confidence; in this sense, confidence provides a form of implicit feedback about performance.

In the late 1970s, Vickers (1978, 1979) developed the idea that confidence provided the basis for adaptive regulation of the decision-making process in the context of his "adaptive module" (see also Baranski & Petrusic, 2003). By generalizing this view to include cognitive performance broadly defined, I suggest that people track systematic changes in their confidence relative to a criterion or target level of confidence. Over time, positive and negative changes (indicating momentary overconfidence and underconfidence, respectively) are accumulated until a criterion level of over- or underconfidence is achieved. For example, as performance begins to decline with increasing SD, the adaptive module will tend to accrue varying magnitudes of "underconfidence." If a criterion amount of underconfidence is achieved, then that would signal that a systematic change in performance has occurred that may, in turn, require behavioral intervention or increased motivation (Horne & Pettitt, 1985; Williams, Lubin, & Goodnow, 1959).

In sum, the correspondence between pretask estimates of performance (i.e., prospective forecasts) and actual performance provides a natural basis for behavioral intervention before a task is undertaken. In this context, it is likely that subjective estimates of fatigue, sleepiness, and motivation provide critical cues that will influence the pretask assessment (Blagrove & Akehurst, 2000; Dorrian et al., 2000). Once engaged in a task, it is likely that direct estimates of task performance will play a more central role in guiding behavior, and I have considered how confidence might provide a basis for the adaptive regulation of behavior. Once the task is complete, the posttask assessments (i.e., retrospective judgments) will play a critical role in feedback learning and thus global calibration of the individual in terms of ability to perform under conditions of fatigue and sleepiness.

Cognitive and Metacognitive Ability During Sleep Deprivation

The results from the perceptual comparison task and the mental addition task showed that participants are able to accurately monitor, respectively, relatively small and relatively large changes in cognitive task performance during SD. From a global perspective, these results suggest that metacognitive abilities are less vulnerable to SD than basic cognitive abilities. Indeed, sleep deprivation is a primitive stressor; thus, on the surface there is some adaptive utility in this relation.

Although it is interesting to consider the present findings from a global perspective, recent advances in the field of neuroscience

allow more specific questions to be asked about cognitive and metacognitive ability during SD. For example, as discussed previously, the large decline in performance on the mental addition task was anticipated because the task involves working memory. To date, considerable research has shown that working memory tasks involve the prefrontal cortex (Chee & Choo, 2004; Funahashi, 2001; Mueller, Machado, & Knight, 2002; Roberts, Robbins, & Weiskrantz, 1998; Rypma, Berger, & D'Esposito, 2002), which is particularly sensitive to the effects of SD (Drummond et al., 1999, 2000; Horne, 1988a, 1993; Muzur, Pace-Schott, & Goldman-Rakic, 2002; Nilsson et al., 2005). On the other hand, there is evidence that the metacognitive ability to self-monitor task performance is based on executive functions (Fernandez-Duque et al., 2000; Hanten, Bartha, & Levin, 2000), also involving the prefrontal cortex (Harrison & Horne, 2000a). One explanation is that certain prefrontal regions, such as those associated with metacognitive functions, may be less sensitive to SD effects than regions associated with the maintenance and manipulation of information in working memory. An alternative explanation is that the executive functions that underlie metacognition are vulnerable to the effects of SD (Harrison & Horne, 2000a; Nilsson et al., 2005), but that there is sufficient compensatory activation in the prefrontal cortex during SD (see, e.g., Chee & Choo, 2004; Drummond & Brown, 2001; Drummond et al., 2000; Drummond, Gillin, & Brown, 2001; Szelenberger, Piotrowski, & Dabrowska, 2005) to preserve higher level monitoring and metacognition. Future research on the neuroscience of SD should explore these avenues with a view toward specifying more precisely the loci of SD effects on cognitive and metacognitive functioning.

In conclusion, there is now converging evidence that the metacognitive ability to self-monitor cognitive performance remains effective for up to 24 hr of SD. Although this result is generally encouraging with respect to self-monitoring ability, it is important to stress two points of caution. First, relatively little is known about the boundary conditions for effective self-monitoring during SD. For example, to date, self-monitoring has been established for simple tasks (e.g., perceptual comparison, mental addition) but remains to be examined in the context of higher level cognitive tasks involving judgment and decision making (see Harrision & Horne, 2000a). Similarly, knowledge is limited concerning the time course of self-monitoring; inevitably, with increasing SD, self-monitoring will likely decline, thus rendering people vulnerable to accidents and injury. Finally, the potential benefits afforded by effective self-monitoring must be tempered by the fact that behavior is ultimately driven by many diverse goals and motivations (e.g., while driving late at night, the desire to reach a destination on time may outweigh the need to stop and rest). Consequently, even perfect self-monitoring is inconsequential when the poor judgments and decisions caused by fatigue are immediate and irreversible (Baranski et al., 1994).

References

Ackerman, P. L., & Wolman, S. D. (2007). Determinants and validity of self-estimates of abilities and self-concept measures. *Journal of Experimental Psychology: Applied, 13,* 57–78.

Angus, R. G., & Heslegrave, R. J. (1985). Effects of sleep loss on sustained cognitive performance during a command and control simulation. *Behavioral Research Methods Instruments & Computers, 17,* 55–67.

Babkoff, H., Caspy, T., & Mikulincer, M. (1991). Subjective sleepiness ratings: The effects of sleep deprivation, circadian rhythmicity and cognitive performance. *Sleep, 14,* 534–539.

Babkoff, H., Caspy, T., Mikulincer, M., & Sing, H. (1991). Monotonic and rhythmic influences: A challenge for sleep deprivation research. *Psychological Bulletin, 109,* 411–428.

Baranski, J. V., Cian, C., Esquivié, D., Pigeau, R. A., & Raphel, C. (1998). Modafinil during 64 hr of sleep deprivation: Dose-related effects on fatigue, alertness, and cognitive performance. *Military Psychology, 10,* 173–193.

Baranski, J. V., Gil, V., McLellan, T., Moroz, D., Buguet, A., & Radomski, M. (2002). Effects of modafinil on cognitive performance during 40 hours of sleep deprivation in a warm environment. *Military Psychology, 14,* 23–48.

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). On the calibration of knowledge and perception. *Canadian Journal of Experimental Psychology, 49,* 397–407.

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 and 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 architectures of the decision–confidence relation. *Canadian Journal of Experimental Psychology, 55,* 195–206.

Baranski, J. V., & Petrusic, W. M. (2003). Adaptive processes in visual perception: Effects of changes in the global difficulty context. *Journal of Experimental Psychology: Human Perception and Performance, 29,* 658–674.

Baranski, J. V., & Pigeau, R. (1997). Self-monitoring cognitive performance during sleep deprivation: Effects of modafinil, d-amphetamine and placebo. *Journal of Sleep Research, 6,* 84–91.

Baranski, J. V., Pigeau, R., & Angus, R. (1994). On the ability to self-monitor cognitive performance during sleep deprivation: A calibration study. *Journal of Sleep Research, 3,* 36–44.

Baranski, J. V., Pigeau, R., Dinich, P., & Jacobs, I. (2004). Effects of modafinil on cognitive and meta-cognitive performance. *Human Psychopharmacology: Clinical and Experimental, 19,* 323–332.

Baranski, J. V., Thompson, M. M., Lichacz, F. M. J., McCann, C., Gil, V., Pasto, L., & Pigeau, R. (2007). Effects of sleep loss on team decision making: Motivational loss or motivational gain? *Human Factors, 49,* 646–660.

Binks, P., Waters, W., & Hurry, M. (1999). Short-term total sleep deprivation does not selectively impair higher cortical functioning. *Sleep, 22,* 328–334.

Björkman, M., Juslin, P., & Winman, A. (1993). Realism of confidence in sensory discrimination: The underconfidence phenomenon. *Perception & Psychophysics, 54,* 75–81.

Blagrove, M., & Akehurst, L. (2000). Effects of sleep loss on confidence–accuracy relationships for reasoning and eyewitness memory. *Journal of Experimental Psychology: Applied, 6,* 59–73.

Blais, A.-R., Thompson, M. M., & Baranski, J. V. (2005). Individual differences in decision processing and confidence judgments in comparative judgment tasks: The role of cognitive styles. *Personality and Individual Differences, 38,* 1701–1713.

Bonnet, M. H., & Arand, D. L. (1995). We are chronically sleep deprived. *Sleep, 18,* 908–911.

Brenner, L., A., Koehler, D. J., Liberman, V., & Tversky, A. (1996). Overconfidence in probability and frequency judgments: A critical ex-

amination. *Organizational Behavior and Human Decision Processes, 65,* 212–219.

Brenner, L. A., Griffin, D., & Koehler, D. (2005). Modeling patterns of probability calibration with random support theory: Diagnosing case-based judgment. *Organizational Behavior and Human Decision Processes, 97,* 64–81.

Buchner, A., Erdfelder, E., & Faul, F. (1997). *How to use G*Power* [Computer software and manual]. Retrieved May 28, 2007, from http://www.psycho.uni-duesseldorf.de/aap/projects/gpower/how_to_use_g-power.html

Chee, M. W. L., & Choo, W. C. (2004). Functional imaging of working memory after 24 hours of total sleep deprivation. *The Journal of Neuroscience, 24,* 4560–4567.

Chua, E. F., Rand-Giovannetti, E., Schacter, D. L., Albert, M. S., & Sperling, R. A. (2004). Dissociating confidence and accuracy: Functional magnetic resonance imaging shows origins of the subjective memory experience. *Journal of Cognitive Neuroscience, 16,* 1131–1142.

Cohen, J. (1988). *Statistical power analyses for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Coren, S. (1997). *Sleep thieves.* New York: Free Press.

Dawson, D., & Fletcher, A. (2001). A quantitative model of work-related fatigue: Background and definition. *Ergonomics, 44,* 144–163.

Dement, W., & Vaughan, C. (1999). *The promise of sleep.* New York: Delacorte Press.

Dinges, D. (1995). An overview of sleepiness and accidents. *Journal of Sleep Research, 4,* 4–11.

Dinges, D. F., & Kribbs, N. B. (1991). Performing while sleepy: Effects of experimentally induced sleepiness. In T. H. Monk (Ed.), *Sleep, sleepiness, and performance* (pp. 97–128). New York: Wiley.

Dorrian, J., Lamond, N., & Dawson, D. (2000). The ability to self-monitor performance when fatigued. *Journal of Sleep Research, 9,* 137–144.

Drummond, S. P., & Brown, G. G. (2001). The effects of total sleep deprivation on cerebral responses to cognitive performance. *Neuropsychopharmacology, 25,* S68–S73.

Drummond, S. P., Brown, G. G., Gillin, J. C., Stricker, J. L., Wong, E. C., & Buxton, R. B. (2000, February 10). Altered brain response to verbal learning followed sleep deprivation. *Nature, 403,* 655–657.

Drummond, S. P., Brown, G. G., Stricker, J. L., Buxton, R. B., Wong, E. C., & Gillin, J. C. (1999). Sleep deprivation-induced reduction in cortical functional response to serial subtraction. *NeuroReport, 10,* 3745–3748.

Drummond, S. P., Gillin, J. C., & Brown, G. G. (2001). Increased cerebral response during a divided attention task following sleep deprivation. *Journal of Sleep Research, 10,* 85–92.

Durmer, J. S., & Dinges, D. F. (2005). Neurocognitive consequences of sleep deprivation. *Seminars in Neurology, 25,* 117–129.

Fernandez-Duque, D., Baird, J. A., & Posner, M. I. (2000). Executive attention and metacognitive regulation. *Consciousness and Cognition, 9,* 288–307.

Ferrell, W. R., & McGoey, P. J. (1980). A model of calibration for subjective probabilities. *Organizational Behavior and Human Performance, 26,* 32–53.

Froberg, J. E. (1977). Twenty-four-hour patterns in human performance, subjective and physiological variables and differences between morning and evening active subjects. *Biological Psychology, 5,* 119–134.

Funahashi, S. (2001). Neuronal mechanisms of executive control by the prefrontal cortex. *Neuroscience Research, 39,* 147–165.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review, 98,* 506–528.

Gillberg, M., Kecklund, G., & Åkerstedt, T. (1994). Relations between performance and subjective sleepiness during a night awake. *Sleep, 17,* 236–241.

Glenville, M., & Broughton, R. (1978). Reliability of the Stanford Sleepiness Scale compared to short duration performance tests and the Wilkinson auditory vigilance task. *Advances in the Biosciences, 21,* 235–244.

Griffin, D., & Buehler, R. (1999). Frequency, probability, and prediction: Easy solutions to cognitive illusions. *Cognitive Psychology, 38,* 48–78.

Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology, 24,* 411–435.

Hanten, G., Bartha, M., & Levin, H. S. (2000). Metacognition following pediatric traumatic brain injury: A preliminary study. *Developmental Neuropsychology, 18,* 383–398.

Harrison, Y., & Horne, J. A. (1998). Sleep deprivation impairs short and novel language tasks having a prefrontal focus. *Journal of Sleep Research, 7,* 95–100.

Harrison, Y., & Horne, J. A. (1999). One night of sleep loss impairs innovative thinking and flexible decision making. *Organizational Behavior and Human Decision Making, 78,* 128–145.

Harrison, Y., & Horne, J. A. (2000a). The impact of sleep deprivation on decision making: A review. *Journal of Experimental Psychology: Applied, 6,* 236–249.

Harrison, Y., & Horne, J. A. (2000b). Sleep loss and temporal memory. *Quarterly Journal of Experimental Psychology, 53,* 271–279.

Harrison, Y., Horne, J. A., & Rothwell, A. (2000). Prefrontal neuropsychological effects of sleep deprivation in young adults—A model for healthy aging? *Sleep, 23,* 1–7.

Harvey, N. (1997). Confidence in judgment. *Trends in Cognitive Science, 1,* 78–82.

Henson, R. N., Rugg, M. D., Shallice, T., & Dolan, R. J. (2000). Confidence in recognition memory for words: Dissociating right prefrontal roles in episodic retrieval. *Journal of Cognitive Neuroscience, 12,* 913–923.

Hoddes, E., Dement, W., & Zarcone, V. (1972). The development and use of the Stanford Sleepiness Scale [Abstract]. *Psychophysiology, 9,* 150.

Hoddes, E., Zarcone, V., Smythe, H., Phillips, R., & Dement, W. (1973). Quantification of sleepiness: A new approach. *Psychophysiology, 10,* 431–436.

Horne, J. A. (1988a). Sleep loss and "divergent" thinking ability. *Sleep, 11,* 528–536.

Horne, J. A. (1988b). *Why we sleep: The functions of sleep in humans and other mammals.* Oxford, England: Oxford University Press.

Horne, J. A. (1993). Human sleep, sleep loss and behavior: Implications for the pre-frontal cortex and psychiatric disorders. *British Journal of Psychiatry, 162,* 413–419.

Horne, J. A., & Pettitt, A. N. (1985). High incentive effects on vigilance performance during 72 hours of total sleep deprivation. *Acta Psychologica, 58,* 133–139.

Horne, J. A., & Reyner, L. A. (1995). Sleep related vehicle accidents. *British Medical Journal, 310,* 565–567.

Johnson, L. C. (1982). Sleep deprivation and performance. In W. Webb (Ed.), *Biological rhythms, sleep, and performance* (pp. 111–141). Chichester, England: Wiley.

Juslin, P. (1993). An explanation of the hard–easy effect in studies of realism of confidence in one's general knowledge. *European Journal of Cognitive Psychology, 5,* 55–71.

Juslin, P. (1994). The overconfidence phenomenon as a consequence of informal experimenter-guided selection of almanac items. *Organizational Behavior and Human Performance, 57,* 226–246.

Juslin, P., & Olsson, H. (1997). Thurstonian and Brunswikian origins of uncertainty in judgment: A sampling model of confidence in sensory discrimination. *Psychological Review, 104,* 344–366.

Keren, G. (1991). Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologica, 77,* 217–273.

Kjellberg, A. (1977). Sleep deprivation and some aspects of performance: I–III. *Waking & Sleeping, 1,* 139–155.

Krueger, G. P. (1989). Sustained work, fatigue, sleep loss and performance: A review of the issues. *Work & Stress, 3,* 129–141.

Leger, D. (1994). The cost of sleep-related accidents: A report for the National Commission on Sleep Disorders Research. *Sleep, 17,* 84–93.

Liberman, V. (2004). Local and global judgments of confidence. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 729–732.

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 and Human Performance, 20,* 159–183.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306–334). Cambridge, England: Cambridge University Press.

Masson, M. E., & Loftus, G. R. (2003). Using confidence intervals for graphically based data interpretation. *Canadian Journal of Experimental Psychology, 57,* 203–220.

Mazzoni, G., & Nelson, T. (1998). *Metacognition and cognitive neuropsychology: Monitoring and control processes.* Mahwah, NJ: Erlbaum.

McClelland, A. G. R., & Bolger, F. (1994). The calibration of subjective probabilities: Theories and models 1980–1994. In G. Wright & P. Ayton (Eds.), *Subjective probability* (pp. 453–482). Chichester, England: Wiley.

Metcalfe, J., & Shimamura, A. P. (1994). *Metacognition: Knowing about knowing.* Cambridge, MA: MIT Press.

Mitler, M., Carskadon, M., Czeisler, C., Dement, W., Dinges, D., & Graeber, R. (1988). Catastrophes, sleep and public policy: Consensus report. *Sleep, 11,* 100–109.

Monk, T. H. (1987). Subjective sleepiness ratings—The underlying circadian mechanisms. *Sleep, 10,* 343–353.

Monk, T. H. (1991). Circadian aspects of subjective sleepiness: A behavioral messenger? In T. H. Monk (Ed.), *Sleep, sleepiness, and performance* (pp. 39–63). New York: Wiley.

Monk, T. H., Leng, V. C., Folkard, S., & Weitzman, E. D. (1983). Circadian rhythms in subjective alertness and core body temperature. *Chronobiologia, 10,* 49–55.

Moses, J., Lubin, A., Naitoh, P., & Johnson, L. C. (1978). Circadian variation in performance, subjective sleepiness, sleep, and oral temperature during an altered sleep–wake schedule. *Biological Psychology, 6,* 301–308.

Mueller, N. G., Machado, L., & Knight, R. T. (2002). Contributions of subregions of the prefrontal cortex to working memory: Evidence from brain lesions in humans. *Journal of Cognitive Neuroscience, 14,* 673–686.

Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology, 12,* 595–600.

Muzur, A., Pace-Schott, E. F., & Goldman-Rakic, P. S. (2002). The prefrontal cortex in sleep. *Trends in Cognitive Sciences, 6,* 475–481.

Nilsson, J. P., Söderström, M., Karlsson, A. U., Lekander, M., Åkerstedt, T., Erixon-Lindroth, N., & Axelsson, J. (2005). Less effective executive functioning after one night's sleep deprivation. *Journal of Sleep Research, 14,* 1–6.

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.

Pilcher, J. J., & Huffcutt, A. I. (1996). Effects of sleep deprivation on performance: A meta-analysis. *Sleep, 19,* 318–326.

Roberts, A. C., Robbins, T. W., & Weiskrantz, L. (1998). *The prefrontal cortex: Executive and cognitive functions.* Oxford, England: Oxford University Press.

Rypma, B., Berger, J. S., & D'Esposito, M. (2002). The influence of working memory demand and subject performance on prefrontal cortical activity. *Journal of Cognitive Neuroscience, 14,* 721–731.

Schneider, S. (1995). Item difficulty, discrimination and the confidence–frequency effect in a categorical judgment task. *Organizational Behavior and Human Decision Processes, 61,* 148–167.

Shimamura, A. (2000). Toward a cognitive neuroscience of metacognition. *Consciousness and Cognition, 9,* 288–307.

Suantak, L., Bolger, F., & Ferrell, W. R. (1996). The "hard–easy effect" in subjective probability calibration. *Organizational Behavior and Human Decision Processes, 67,* 201–221.

Szelenberger, W., Piotrowski, T., & Dabrowska, A. J. (2005). Increased prefrontal event-related current density after sleep deprivation. *Acta Neurobiologiae Experimentalis, 65,* 19–28.

Thomas, M., Sing, H., Belenky, G., Holcomb, H., Mayberg, H., Dannals, R., et al. (2002). Neural basis of alertness and cognitive performance impairments during sleepiness: I. Effects of 24h of sleep deprivation on waking human regional brain activity. *Journal of Sleep Research, 9,* 335–352.

Vickers, D. (1978). An adaptive module for simple judgments. In J. Requin (Ed.), *Attention and performance VII* (pp. 599–618). Hillsdale, NJ: Erlbaum.

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

Webb, W. B. (1995). The cost of sleep related accidents: A reanalysis. *Sleep, 18,* 276–280.

Wilkinson, R. T. (1965). Sleep deprivation. In O. Edholm & A. Bacharach (Eds.), *The physiology of human survival* (pp. 399–430). New York: Academic Press.

Wilkinson, R. T. (1969). Sleep deprivation: Performance tests for partial and selected sleep deprivation. In L. E. Abt & B. F. Riess (Eds.), *Progress in clinical psychology* (pp. 28–43). New York: Grune & Stratton.

Williams, H. L., Lubin, A., & Goodnow, J. J. (1959). Impaired performance with acute sleep loss. *Psychological Monographs: General and Applied, 73,* 1–26.

Wimmer, F., Hoffmann, R. F., Bonato, R. A., & Moffitt, A. R. (1992). The effects of sleep deprivation on divergent thinking and attention processes. *Journal of Sleep Research, 1,* 223–230.

Yaniv, I., Yates, J. F., & Smith, J. E. K. (1991). Measures of discrimination skill in probabilistic judgment. *Psychological Bulletin, 110,* 611–617.

Yates, J. F. (1982). External correspondence: Decompositions of the mean probability score. *Organizational Behavior and Human Decision Processes, 30,* 132–156.

Yates, J. F. (1990). *Judgment and decision making.* Englewood Cliffs, NJ: Prentice Hall.

Received December 1, 2005
Revision received June 4, 2007
Accepted June 21, 2007 ■