Image Cover Sheet

CLASSIFICATION	SYSTEM NUMBER
UNCLASSIFIED	389685

TITLE

ON THE ABILITY TO SELF-MONITOR COGNITIVE PERFORMANCE DURING SLEEP-DEPRIVATION: A CALIBRATION STUDY

System Number:

Patron Number:

Requester:

Notes:

DSIS Use only:

Deliver to: FF

J. Sleep Res. (1994) 3, 36–44

On the ability to self-monitor cognitive performance during sleep deprivation: a calibration study

JOSEPH V. BARANSKI, ROSS A. PIGEAU and ROBERT G. ANGUS
Defence and Civil Institute of Environmental Medicine, Human Factors Division, North York, Ontario, Canada

Accepted in revised form 26 October 1993; manuscript received 21 June 1993

SUMMARY The antagonistic effects of extensive sleep deprivation (SD) on human cognitive performance are well documented. However, one aspect of human performance that has not been investigated with respect to its susceptibility to SD is the 'meta-cognitive' ability to self-monitor overt performance. In the present study, 16 male subjects participated in an experiment requiring sustained cognitive work during a three day period. One of the cognitive tasks required the mental addition of rapidly presented numbers. On each trial, subjects reported the sum and then provided a subjective confidence rating to indicate the degree of certainty in their response. As expected, performance on the sequential addition task deteriorated with increasing fatigue and returned to baseline following a recovery sleep. However, *calibration* analyses, which quantify a number of properties of the relationship between subjective and overt performance, revealed that the correlation between confidence and performance (calibration), the ability to differentiate correct from incorrect judgments (resolution), and validity of subjective 'certainty', were all unaffected by SD. Hence, in the absence of external feedback from the environment, people have access to fairly reliable internal feedback about their performance during periods of sustained and vigilant cognitive activity.

KEYWORDS calibration, confidence, meta-cognition, performance, sleep deprivation

INTRODUCTION

The antagonistic effects of extensive sleep deprivation (SD) on human cognitive performance are well documented (for reviews see Wilkinson 1965, 1969; Naitoh and Townsend 1970; Naitoh 1976; Kjellberg 1977; Johnson 1979, 1982; Horne 1988; Naitoh and Angus 1989; Babkoff *et al.* 1991b). However, one aspect of human performance that has not been investigated with respect to its susceptibility to SD is the 'meta-cognitive' ability to self-monitor overt performance.

Research on the ability to self-monitor cognitive performance can be traced back to Peirce and Jastrow's (1884) classic psychophysical study on the comparison of successively lifted weights. Experimenting on themselves, Peirce and Jastrow selected, on each of several thousand trials, one of the two 'pressures' as the greater and then reported a subjective *confidence* rating (on a four-point scale) to indicate the degree of certainty in their judgment. Among other landmark findings, their experiments revealed a strong and positive correspondence between the degree of subjective confidence and the proportion of correct judgments.

Current research on the relationship between subjective and overt performance is proceeding along several avenues. One line of research, continuing in the classical psychophysical tradition (e.g. Henmon 1911; Johnson 1939; Festinger 1943), is studying the many interesting and complex interactions among confidence, judgment accuracy, and response time (RT), in the context of developing and evaluating broader theories of psychophysical judgment and discrimination (e.g. Vickers 1979; Vickers and Packer 1982; Heath 1984; Vickers *et al.* 1985; Petrusic and Baranski 1989a,b; Link 1992; Petrusic 1992). Another line of research is more specifically concerned with the relationship between confidence and performance, focusing on systematic properties, various factors that might affect or improve the relationship, and, more recently, specific theories (e.g. Ferrell and McGoey 1980; Gigerenzer *et al.* 1991; Griffin and Tversky 1992). The latter line of investigation falls under the general heading of *calibration* research (for

Correspondence: Dr. Joseph V. Baranski DCIEM, Human Factors Division 1133 Sheppard Ave. West P.O. Box 2000 North York, Ontario, Canada M3M 3B9.

reviews see Lichtenstein *et al.* 1982; Yates 1990; Keren 1991).

In a typical calibration study, the subject is required to give repeated observations on some task and to provide 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 subject cannot accurately self-monitor performance and thus is said to have poor calibration.

To date, most calibration studies have required subjects to report their confidence in terms of a 'subjective probability' (de Finetti 1937; Savage 1954; Phillips 1973; Wallsten and Budescu 1983), where, for example, a subjective probability of 0.5 (i.e. 50% correct) denotes a guessing response (in a two-alternative task) and a subjective probability of 1.0 (i.e. 100% correct) denotes 'certainty'. In addition, there are well-developed analytical procedures for assessing different aspects of the relationship between overt and subjective probabilities (see Murphy 1973; Yates 1982; Yaniv *et al.* 1991; Baranski and Petrusic in press). Alternatively, a number of recent studies, most notably those interested in the calibration of reading comprehension (e.g. Glenberg and Epstein, 1985, 1987; Glenberg *et al.* 1987; Morris 1990; Weaver 1990) and the accuracy of eyewitness memory/identification (e.g. Wells *et al.* 1979; Smith *et al.* 1989; Perfect *et al.* 1993), have opted for an ordinal confidence scale (e.g. 1–6) and the primary index of calibration was the correlation between confidence and overt performance.

The present study employed correlational calibration analyses (and others to be described) to investigate the effect of extensive SD on the ability to self-monitor cognitive performance. Subjects performed a number of cognitive tasks during the experiment but we focus on confidence data collected following each trial of a serial addition task with six *a priori* defined levels of difficulty. Although the study was concerned primarily with the effects of sustained operation conditions on human performance in a simulated military command and control environment, the results of this research can apply, more generally, to a variety of situations requiring sustained and vigilant cognitive activity (e.g. search and rescue, fire fighting, medical and civil emergencies, lengthy work cycles, etc. . .).

METHOD

Subjects

Sixteen male Canadian Forces volunteers each served for approximately 100 h of experimental participation. Subjects were informed about the purpose of the experiment, the procedures to be employed, and understood that they were free to withdraw from the experiment at any time.

Materials

The laboratory was effectively isolated from the normal activities of the research institute and contained facilities for accommodating the needs of both subjects and experimenters. Subjects worked alone in 3×4 m experimental rooms. Each room was equipped with a DEC VT100 video display terminal, table, chair, and a desk lamp. All cognitive tasks were controlled by a DEC VAX 7/85 computer and were displayed on the subjects' terminals. Subjects responded by keying their answers on the terminal keyboard. Closed-circuit televisions were used to monitor the subjects and slave monitors displayed the subjects' responses to the experimenters. Hence by monitoring both the subjects and their responses the experimenters were able to determine when subjects fell asleep and immediately awaken them.

Procedure

Subjects participated in groups of four, but worked independently of each other and at their own pace. They arrived at 10.00 hours on Monday and remained in the laboratory until 15.00 hours on Friday. All time cues were removed and interpersonal communication with the laboratory staff was kept to a minimum. Upon arrival, subjects were briefed on the experimental protocol and were given extensive practice on the battery of cognitive tasks to be used in the experiment. Training continued until 20.00 hours on Monday, after which the subjects relaxed, watched a movie, and retired for 8 h of baseline sleep at 22.00 hours. They were awakened at 06.00 hours on Tuesday morning and immediately began the experiment. On Thursday morning at 04.00 hours they were given a 2 h nap (unannounced), after which they immediately continued the experiment until 20.00 hours on Thursday night. They then relaxed and went to bed at 22.00 hours for 8 h of recovery sleep. They were woken at 06.00 hours on Friday and worked until 15.00 hours, at which time the experiment ended.

Throughout the SD period the subjects performed continuous cognitive work for 1 h and 45 min followed by a 15 min break. During the breaks, subjects consumed food, used the restroom facilities, watched movies, and conversed with the other subjects. EEG, ECG, and surface body temperatures were recorded throughout the study (these data are not reported here). The cognitive performance tasks included complex iterative subtraction, logical reasoning, encoding/decoding, short-term memory (digit span), paired-associate learning/recall, map plotting, message processing, auditory vigilance, serial four-choice reaction time, and serial addition. Some tasks (e.g. self-report scales) were presented in each 2-h session while

other tasks (e.g. serial addition) appeared only once in each 6-h block. The order of presentation of the tasks was held constant within the 6-h blocks and each task was timed so that subjects could work at their own pace but complete the test sessions at approximately the same time (and thus spend their breaks together).

Confidence ratings were collected during each trial of the serial addition task. Hence, only results pertaining to that task will be reported here. For illustrative results on the other tasks we refer the reader to previous DCIEM research (e.g. Angus and Heslegrave 1985; Heslegrave and Angus 1985; Pigeau *et al.* 1987a 1987b; Naitoh and Angus 1989; Angus *et al.* 1992).

The sequential addition task required subjects to mentally add a sequence of 8 numbers which were presented on the computer monitor at a rate of one number every 1.25 s. The sequence was terminated by the presentation of a visual prompt ($\Rightarrow$) at which time subjects typed in their response and then pressed the 'Enter' key. RTs were recorded from the presentation of the prompt until the 'Enter' key was pressed. Following each trial, the subjects were prompted to key in a confidence rating, from 1 to 6, to reflect the degree of certainty in their response. A rating of 1 denoted a complete lack of certainty in the response ('Not Sure') and a rating of 6 denoted subjective certainty ('Sure'). Confidence ratings between 2 and 5 were to correspond to increasing certainty in the correctness of the response. Subjects depressed the 'Enter' key to record their confidence ratings and to proceed to the next trial.

The task included six *a priori* defined levels of trial difficulty, where the levels of difficulty were monotonically related to the magnitudes of the numbers to be added (see Ashcraft 1992). Specifically, on each trial, 8 numbers were randomly presented from one of the following sets which define the six levels of difficulty:

Level 1: 1,2.
Level 2: 2,3,4,5,6.
Level 3: 4,5,6,7,8.
Level 4: 6,7,8,9,12.
Level 5: 8,9,12,13,14.
Level 6: 12,13,14,15,16.

Hence, for example, a trial comprising difficulty level 1 would involve a random presentation of the numbers 1 and 2 (e.g. 1, 1, 2, 1, 1, 1, 2, 2). The levels of difficulty were presented randomly, without replacement, from the pool of six. This ensured that, overall, each level appeared equally often. Subjects performed 10 min of the serial addition task in each of the 11 six-hour blocks in the study. On average, each subject provided approximately 50 observations per block. Subjects 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.

RESULTS

The results are presented in three sections. The fir examines the effects of trial difficulty and experiment block (i.e. fatigue, nap, and recovery sleep) on performanc in the serial addition task. Section two provides th calibration analyses, focusing on the effects of fatigue on th ability to self-monitor performance. Finally, section thre provides additional analyses that highlight the similaritie between the present data and those of previous studies.

Global analyses

Separate repeated measures analyses of varianc (ANOVAs) were conducted on the mean RT, probabilit correct (arcsine transformed), and confidence ratin measures of performance. There were two within-subje factors: trial difficulty (6 levels) and experimental block (1 levels)[1]. Since the difficulty level × block interaction was n reliable for any of the dependent measures, we report onl the main effects. All 16 subjects completed the experimer and the overall proportion of correct responses in the tas was 0.739 ($N = 8857$).

Effects of trial difficulty. Figure 1 provides plots of th proportion correct, subjective confidence and mean RT fc each level of trial difficulty, after combining the data ov subjects and blocks. The plots clearly show the effectivene of the *a priori* difficulty manipulation; RTs increas ($F_{(5,75)} = 30.70$, $P < 0.0001$) and proportion corre ($F_{(5,75)} = 19.70$, $P < 0.0001$) and subjective confiden ($F_{(5,75)} = 20.63$, $P < 0.0001$) decrease as trial difficul increases.

Effects of fatigue, napping, and recovery sleep. Figure provides plots of the proportion correct, subjective co fidence and mean RT for each of the 11 experiment blocks, after combining the data over subjects and difficult levels. As expected, performance declined with increasin fatigue, a 2-h nap provided some restoration of perfo mance, and an 8-h sleep restored performance to baselin Importantly, the subjective confidence measure is quit sensitive to overt performance and, overall, displays pattern that is very similar to that observed for th proportion correct measure. Finally, RTs show a gener trend that, likewise, corroborates the finding of declinin performance with fatigue and recovery following sleep.

The main effect of block was highly reliable for respon accuracy ($F_{(10,150)} = 3.67$, $P < 0.01$), RT ($F_{(10,150)} = 6.1$ $P < 0.0001$), and confidence ($F_{(10,150)} = 7.00$, $P < 0.0001$

[1] All significance levels are based on the Greenhouse–Geiss adjusted degrees of freedom. However, the degrees of freedo reported in the text are based on the design. In addition, o subject did not provide data for block 10 because he had to use th washroom. In each analysis, the data point was estimated accordi to the recommendation of Myers and Well (1991).

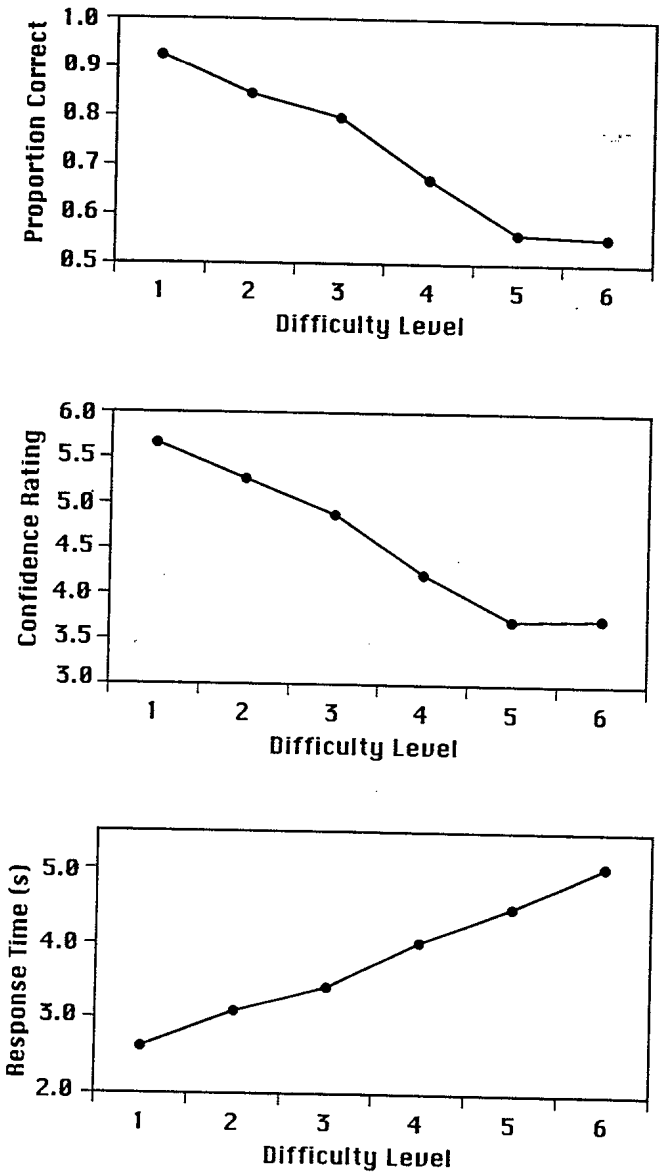

Figure 1. Mean proportion correct, confidence rating, and RT associated with each level of trial difficulty. Each point is based on approximately 1450 observations.

confirming that performance differed over the 11 blocks. Trend analyses revealed that the linear effect of block was not reliable for any of the dependent measures. However, the quadratic (RT: $F_{(1,15)} = 15.30$, $P < 0.002$; P(c): $F_{(1,15)} = 15.60$, $P < 0.002$; confidence: $F_{(1,15)} = 67.90$, $P < 0.0001$) and cubic (RT: $F_{(1,15)} = 7.46$, $P < 0.016$; P(c): $F_{(1,15)} = 10.84$, $P < 0.005$; confidence: $F_{(1,15)} = 12.75$, $P < 0.003$) components were highly reliable, confirming the trends in the data evident in Fig. 2.

In sum, these global analyses confirm that performance in the serial addition task was clearly sensitive to the effects of trial difficulty and SD. Overall, confidence ratings closely track the proportion correct measure suggesting that subjects are aware of their level of performance over the course of the experiment. However, more detailed aspects of self-monitoring, such as the linear correspondence between confidence and performance and the ability to differentiate correct from incorrect judgments on a trial-by-trial basis, cannot be quantified by such global analyses. The following section examines these specific aspects of self-monitoring.

Calibration analyses

Fatigue and the relationship between confidence and performance. Figure 3 provides the 'calibration curve' for all data collected in the study. The calibration curve is obtained by plotting the proportion of correct responses associated with each confidence level. The plot shows (a) that subjects are very likely to be incorrect when they think they are (i.e. at confidence level 1), (b) that subjects are very likely to be correct when they are 'certain' (i.e. at confidence level 6), and, (c) that performance monotonically increases through the intermediate confidence levels; overall, these subjects appear to be 'well-calibrated' in the process of mental addition.

We computed the Pearson product-moment correlation (**r**) between confidence (1–6) and performance (0 for incorrect, 1 for correct) for each subject in each block.[2] The mean of individual subject correlations was 0.64 ($P < 0.0001$), reflecting good calibration. However, the main effect of block was not reliable ($F < 1.0$), and a trend analysis revealed no reliable trends in the data.[3] The correlations between confidence and accuracy based on the overall data are plotted for each block in Fig. 4.

The effect of fatigue on confidence 'resolution'. The correlation between confidence and performance provides one component of calibration analyses with ordinal confidence scales. Another important property of self-monitoring is the degree to which subjects can assign a maximally different

[2] The primary statistics for calibration analyses with ordinal confidence scales have been the Pearson product-moment correlation **r** (e.g. Glenberg and Epstein 1985; Glenberg *et al.* 1987; Morris 1990; Weaver 1990) and the non-parametric gamma correlation, **G** (see Nelson 1984; Stock *et al.* 1992; Perfect *et al.* 1993; Schneider and Laurion 1993). We performed all analyses using both of these measures and the interpretation of the results was the same (see Weaver 1990). Hence, for convenience in data presentation, only Pearson correlations will be reported in this paper. In addition, for all ANOVAs involving correlations, the Pearson coefficients were normalized using the Fisher **r** to **z** transformation. The Fisher values were then transformed back and are reported, for convenience, as Pearson coefficients in the text and in the figures.

[3] Two subjects made no errors and always reported certainty in a few of their blocks. Hence a correlation could not be computed because there was no variability in their confidence ratings or their performance. In one ANOVA we excluded these subjects and in another ANOVA we entered a correlation of 1.0 for the missing cells (i.e. 'perfect calibration'). The interpretation of the ANOVAs was the same in each case.

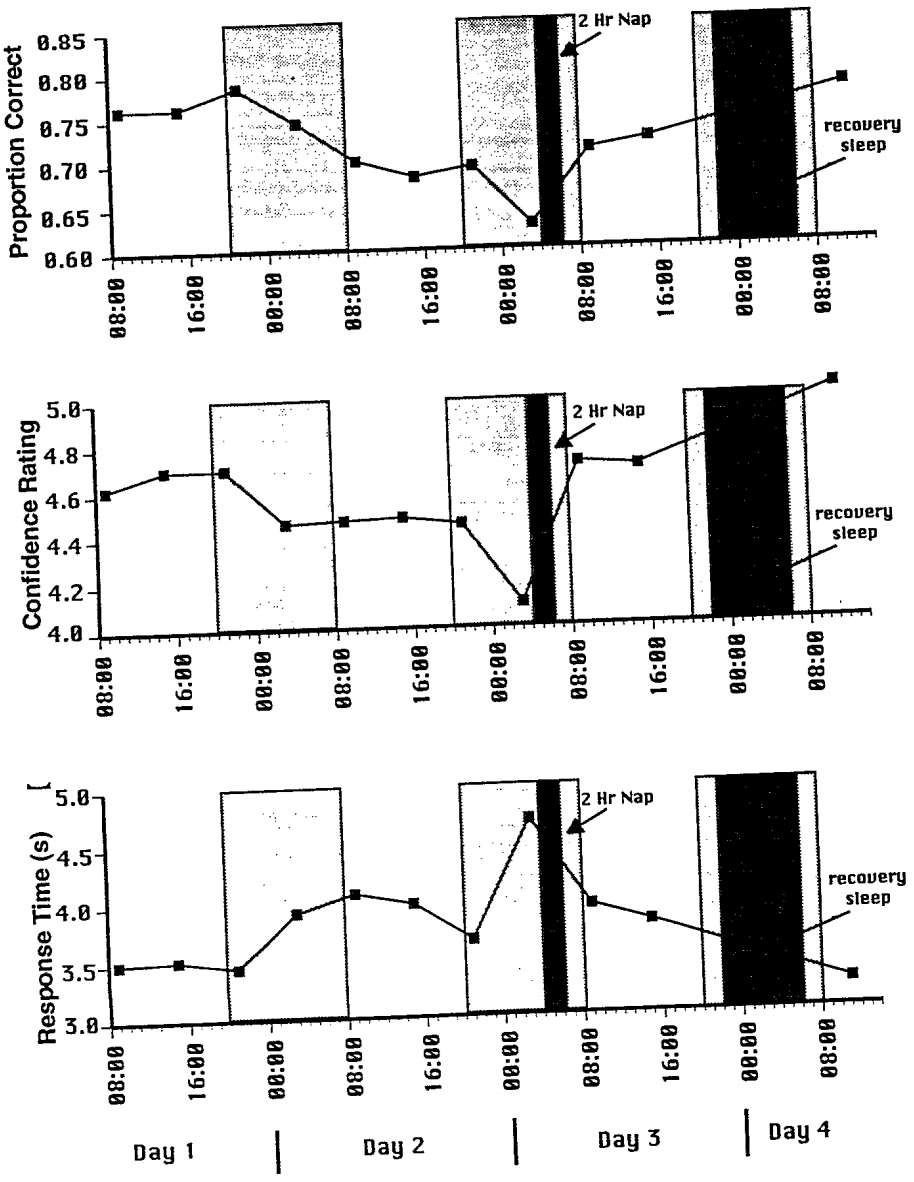

Figure 2. Mean proportion correct, confidence rating, and RT associate each of the 11 experimental blocks. shaded areas denote evening and ea morning periods and dark shaded a denote time spent asleep. Each poi based on approximately 800 observ

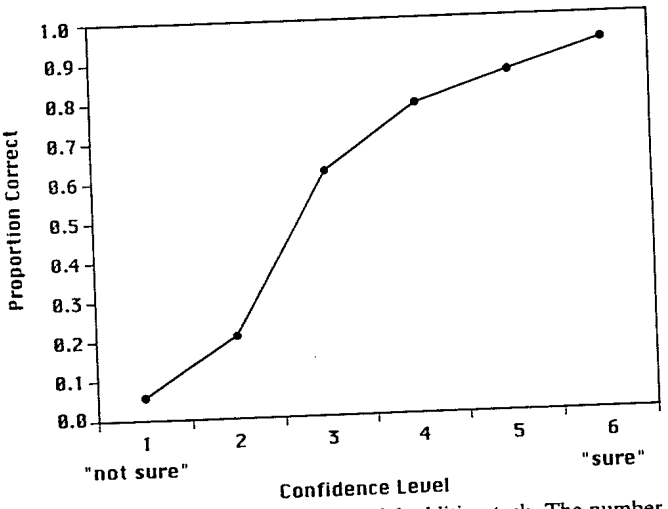

Figure 3. Calibration curve for the serial addition task. The number of observations associated with each confidence level are 1289, 309, 414, 954, 1659, and 4232 for levels 1–6, respectively.

proportion correct to the various confidence categ the degree to which subjects can use their confide to differentiate correct from incorrect response context of subjective probability analyses, this skil as resolution (Murphy 1973) or *discrimination* (Y Yaniv *et al*. 1991). With respect to the ordinal scale employed in the present study, resoluti denoted by the *slope* of the calibration curve.

We determined the best fitting regression li calibration curve of each subject in each block. T the regression line was used as a dependent var ANOVA with the 11 experimental blocks as subjects factor. The ANOVA confirmed that were significantly different from zero ($P < 0.00$ ver, the effect of block was not reliable ($F <$ trend analysis revealed no reliable trends in Hence, neither calibration (as defined by the between confidence and performance) nor re

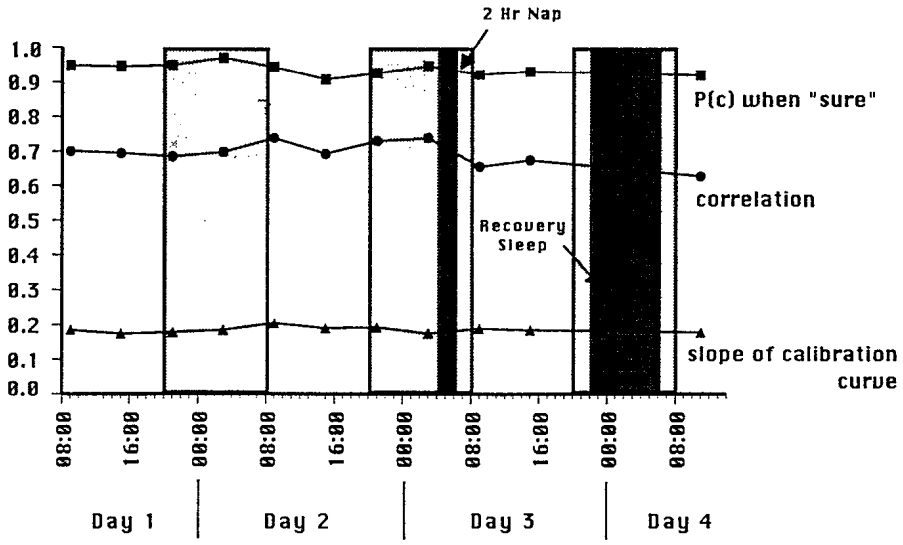

Figure 4. Proportion correct associated with subjective 'certainty' (top), correlation between confidence and performance (middle) and slope of the calibration curve (bottom), for each of the 11 blocks. Lightly shaded areas denote evening and early morning periods and dark shaded areas denote time spent asleep. Points are based on the overall data.

defined above) were affected by SD. The regression slopes for each block are plotted in Fig. 4.

The effect of fatigue on the validity of subjective 'certainty'. An important adjunct component of calibration analyses centres on the level of performance associated with subjective 'certainty'. Ideally, people should always be correct when they are certain that they are but factors such as task or trial difficulty can lead to substantial departures from this ideal (Fischhoff *et al.* 1977). In line with intuition, if the proportion correct associated with subjective certainty is very high then there exists a high degree of validity associated with that confidence level. Conversely, progressive departure from optimal performance denotes increasing *over-confidence* in the judgment of certainty.

In the present task, the overall proportion correct associated with the 'certain' confidence level was 0.937 ($N = 4232$). In other words, there was only a slight degree of over-confidence associated with the subjective certainty report (cf. Lichtenstein *et al.* 1982; Glenberg and Epstein 1985; Paese and Sniezek 1991, for demonstrations of substantial overconfidence in other cognitive tasks). Importantly, as is evident in Fig. 4, this degree of over-confidence was not affected by fatigue, the nap, or the recovery sleep.

We determined the proportion correct associated with the 'certain' confidence level for each subject in each block.[4] An ANOVA conducted on these (arcsine transformed) proportions confirmed that the effect of block was not reliable and a trend analysis revealed no reliable trends in the data.

[4] The data of two subjects was not included in this ANOVA because they did not use the 'certain' confidence level in each block.

Additional analyses

The present findings suggest that various aspects of the ability to self-monitor cognitive performance are not antagonistically affected by SD. Although the generalizability of the present findings to other cognitive tasks must, of course, await future studies, some indication may be gained by demonstrating that the present data display other properties of confidence calibration that are well documented in a broad range of performance tasks.

The calibration 'difficulty effect'. Numerous studies have found that people have a tendency to be over-confident on difficult tasks (i.e. when the proportion of correct responses is low) and under-confident on relatively easy tasks. The phenomenon is referred to as the calibration *difficulty effect* and has been observed in virtually all studies that have investigated wide ranges in task difficulty (Lichtenstein and Fischhoff 1977; Griffin and Tversky 1992). Typically, in most judgment domains, 'calibration improves up to about 80% correct, and then becomes worse' (Lichtenstein and Fischhoff 1977, p.180).

We obtained a calibration curve for each subject for difficulty levels 1–2, 3–4, and 5–6 (the 6 levels of trial difficulty were combined into three because several subjects did not use the lower confidence ratings for difficulty level 1). The slope of the calibration curve and the correlation between confidence and the proportion of correct responses associated with each confidence level were used as dependent measures in an ANOVA with the three levels of trial difficulty as a within-subjects factor. As is evident in Table 1, the correlation and the slope of the calibration curve are highest for the intermediate difficulty level (approximately the optimal proportion correct level discussed by Lichtenstein and Fischhoff 1977 for general

Table 1 Proportion correct, correlation, and the slope of the calibration curve for trial difficulty levels 1–2, 3–4, and 5–6.

Level	*P(c)*	*Correlation*	*Slope*
1–2	0.900	0.792	0.110
3–4	0.748	0.911	0.162
5–6	0.562	0.764	0.133

Note Data are based on the mean of individual subject scores.

knowledge questions) and drop off for the harder and easier levels where accuracy is either very high (Level 1) or very low (Level 3). Trend analyses confirmed that the linear effect of difficulty level was not reliable for either the slope ($F < 1.65$) or the correlation ($F < 1.0$). However, the quadratic component was highly reliable for both the slope ($F_{(1,15)} = 10.11$, $P < 0.006$) and the correlation ($F_{(1,15)} = 6.15$, $P < 0.025$), confirming a calibration difficulty effect with the present ordinal confidence scale.

The RT 'difficulty effect'. Figure 5 provides plots of the mean RT associated with each level of confidence for the three levels of difficulty described above. The plots show that (a) RTs are inversely related to confidence (see e.g. Henmon 1911; Volkmann 1934; Vickers *et al.* 1985, for demonstrations with perceptual judgments), and (b) the functions relating the two measures are ordered with respect to the level of difficulty of the judgments; i.e. there is an RT 'difficulty effect'. Importantly, these data provide a replication of studies investigating the properties of RT and calibration in perceptual (Baranski and Petrusic in press) and memory-based intellectual knowledge (Baranski and Petrusic unpubl. data) judgment tasks (see Baranski and Petrusic 1992 for additional similarities between perception and memory). As discussed by Baranski and Petrusic (in press), the RT difficulty effect is important in ruling out the possibility that confidence is *determined* by scaling the duration of the decision or judgment process, i.e. a specific judgment time is not associated with a specific confidence level.

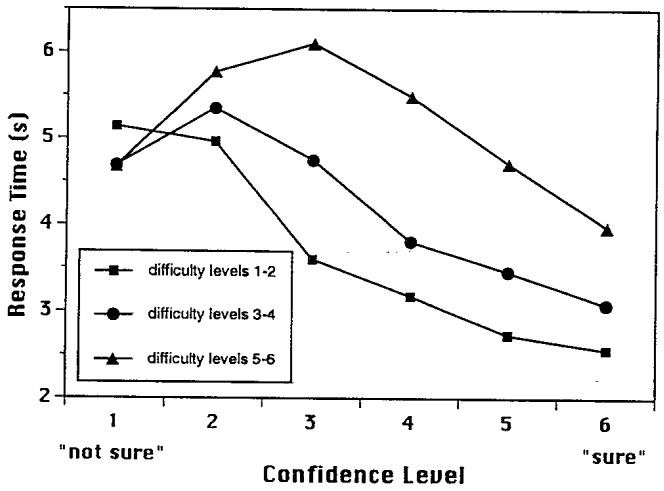

Figure 5. Mean RTs associated with each level of confidence for difficulty levels 1–2, 3–4, and 5–6.

An ANOVA was conducted with mean RT as the dependent measure and trial difficulty (3 levels) and confidence level (3 levels)[5] as within subject factors. The main effects of difficulty level ($F_{(2,28)} = 19.85$, $P < 0.0001$) and confidence level ($F_{(2,28)} = 7.95$, $P < 0.006$) were highly reliable, confirming the ordering of the curves according to difficulty level and the inverse relationship between confidence and RT, respectively. The interaction between difficulty level and confidence level was reliable by a conventional test ($F_{(4,56)} = 2.77$, $P < 0.04$) but not with the Greenhouse–Geisser degrees of freedom ($0.05 < P < 0.08$). The latter effect is likely due to the relatively quick 'not sure' (i.e. confidence level 1) responses made for the higher levels of difficulty (see Glucksberg and McCloskey 1981).

DISCUSSION

The results of the present study replicate and extend the well known finding of declining cognitive performance over extensive periods of SD. In contrast, the 'meta-cognitive' ability to self-monitor performance is apparently not affected by SD. Specifically, global analyses showed that subjective confidence ratings closely track overt performance and calibration analyses revealed that the correlation between confidence and performance (i.e. calibration), the ability to differentiate correct from incorrect judgments (i.e. resolution), and the validity of subjective certainty, are not affected by SD. Hence, during SD, subjective reports can not only quantify fatigue (e.g. Harris *et al.* 1971), sleepiness (e.g. Hoddes *et al.* 1973; Herscovitch and Broughton 1981; Dinges 1989; Babkoff *et al.* 1991a), and mood (e.g. Johnson and Naitoh 1974), perhaps more importantly, they can provide a basis for the accurate self-monitoring of overt performance.

It is interesting to note that calibration is affected by performance variations induced by different levels of problem difficulty (i.e. the calibration 'difficulty effect') but not by comparable performance variations induced by fatigue. This finding suggests a 'context-free' process of confidence estimation in which subtle limitations in the mental algorithm used to scale trial-by-trial variations in problem difficulty (e.g. Ferrell and McGoey 1980; Gigerenzer *et al.* 1991; Griffin and Tversky 1992), are not accentuated by a global increase in task difficulty induced by fatigue.

To elaborate, we first note that considerable research has implicated attentional deficits as the basis for cognitive performance degradations observed during SD (e.g. Norton 1970; Kjellberg 1977; Sanders and Reitsma 1982;

[5] The 6 confidence levels were combined into 3 because each subject did not use every confidence level. Even so, the data for one subject was not included because they did not use confidence level 5 or 6 for the easiest difficulty level.

Wimmer *et al*. 1992). Secondly, in the context of the present task, current theories of complex mental addition (Ashcraft 1992; Dehaene 1992; Gallistel and Gelman 1992) assume a number of component processes (e.g. encoding, fact retrieval, short-term memory maintenance, carrying operations, . . .) and detailed theoretical analyses of these components have recently been presented (e.g. Frensch and Geary 1993). Whereas problem difficulty is a function of the demands of each of the above-mentioned component processes, the effect of fatigue might be to selectively disrupt those component processes that require sustained attentional resources (e.g. encoding and maintenance). Indeed, because most adults are highly skilled in the process of mental addition, we may be particularly sensitive to extraordinary perturbations in the processes that regulate overt performance. Accordingly, future research on calibration and SD should investigate more complex, 'higher' level, cognitive tasks and/or tasks in which there does not exist a high level of expertise.

In conclusion, the present findings suggest that, in the absence of *external feedback* from the environment, people have access to fairly accurate *internal feedback* about their performance during work cycles requiring prolonged and vigilant cognitive activity. We should keep in mind, however, that the obvious benefits of good calibration must always be tempered by the fact that even perfect calibration is useless when the consequences of poor judgments caused by fatigue, or any other factor, are immediate and irreversible.

ACKNOWLEDGEMENTS

We would like to thank Jim A. Horne, William M. Petrusic, Megan M. Thompson and two anonymous reviewers for their very helpful comments and Andrea Hawton for assistance with the figures.

REFERENCES

Angus, R.G. and Heslegrave, R.J. Effects of sleep loss on sustained cognitive performance during a command and control simulation. *Behav. Res. Meth. Inst. & Comp*., 1985, 17: 55–67.

Angus, R.G., Pigeau, R.A. and Heslegrave, R.J. Sustained operations studies: From the field to the laboratory. In: C.Stampi (Ed.) *Why We Nap*. Birkhauser: Boston, 1992: 217–241.

Ashcraft, M.H. Cognitive arithmetic: A review of the data and theory. *Cognit*., 1992, 44: 75–106.

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

Babkoff, H., Caspy, T., Mikulincer, M. and Sing, H. Monotonic and rhythmic influences: A challenge for sleep deprivation research. *Psychol. Bull*., 1991b, 109: 411–428.

Baranski, J.V. and Petrusic, W.M. The discriminability of remembered magnitudes. *Mem. & Cognit*., 1992, 20: 254–270.

Baranski, J.V. and Petrusic, W.M. The calibration and resolution of confidence in perceptual judgments. *Perc. & Psychophys*., in press.

de Finetti, B. La prevision: Ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincare*, 1937, 7: 1–68. English translation in H.E. Kyburg, Jr. and H.E. Smokler (Eds) *Studies in Subjective Probability*. Wiley, New York, 1964.

Dehaene, S. Varieties of numerical abilities. *Cognit*., 1992, 44: 1–42.

Dinges, D.F. The nature of sleepiness: Causes, context and consequences. In: A.J. Stunkard and A. Baum (Eds) *Perspectives in Behavioral Medicine*. Erlbaum, Hillsdale, 1989: 147–179.

Ferrell, W.R. and McGoey, P.J. A model of calibration for subjective probabilities. *Org. Behav. Hum. Perf*., 1980, 26: 32–53.

Festinger, L. Studies in decision: I. Decision-time, relative frequency of judgment and subjective confidence as related to physical stimulus difference. *J. Exp. Psychol*., 1943, 32: 291–306.

Fischhoff, B., Slovic, P. and Lichtenstein, S. Knowing with certainty: The appropriateness of extreme confidence. *J. Exp. Psychol.: Hum. Perc. & Perf*., 1977, 3: 552–564.

Frensch, P.A. and Geary, D.C. Effects of practice on component processes in complex mental addition. *J. Exp. Psychol.: Learn., Mem., & Cognit*., 1993, 19:433–456.

Gallistel, C.R. and Gelman, R. Preverbal and verbal counting and computation. *Cognit*., 1992, 44: 43–74.

Gigerenzer, G., Hoffrage, U. and Kleinbölting, H. Probabilistic mental models: A Brunswikian theory of confidence. *Psychol. Rev*., 1991, 98: 506–528.

Glenberg, A.M. and Epstein, W. Calibration of comprehension. *J. Exp. Psychol.: Learn., Mem. & Cognit*., 1985, 11: 702–718.

Glenberg, A.M. and Epstein, W. Inexpert calibration of comprehension. *Mem. & Cognit*., 1987, 15: 84–93.

Glenberg, A.M., Sanocki, T., Epstein, W. and Morris, C. Enhancing calibration of comprehension. *J. Exp. Psychol.: Gen*., 1987, 116: 119–136.

Glucksberg, S. and McCloskey, M. Decisions about ignorance: Knowing that you don't know. *J. Exp. Psychol.: Hum. Learn. & Mem*., 1981, 7: 311–325.

Griffin, D. and Tversky, A. The weighing of evidence and the determinants of confidence. *Cog. Psychol*., 1992, 24: 411–435.

Harris, D.A., Pegram, G.V. and Hartman, B.O. Performance and fatigue in experimental double-crew transport missions. *Aviat. Space & Env. Med*., 1971, 24: 980–986.

Heath, R.A. Random-walk and accumulator models of psychophysical discrimination: A critical evaluation. *Percept*., 1984, 13: 57–65.

Henmon, V.A.C. The relation of the time of a judgment to its accuracy. *Psychol. Rev*., 1911, 18: 186–201.

Herscovitch, J. and Broughton, R. Sensitivity of the Stanford sleepiness scale to the effects of cumulative partial sleep deprivation and recovery oversleeping. *Sleep*, 1981, 4: 83–92.

Heslegrave, R.J. and Angus, R.G. The effects of task duration and work-session location on performance degradation induced by sleep loss and sustained cognitive work. *Behav. Res. Meth. Inst. & Comp*., 1985, 17: 592–603.

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

Horne, J.A. *Why We Sleep: The functions of sleep in humans and other mammals*. Oxford University Press, Oxford, 1988.

Johnson, D.M. Confidence and speed in the two-category judgment. *Arch. Psychol*., 1939, 34: 1–53.

Johnson, L.C. Sleep disturbances and performance. In: A.N. Nicholson (Ed.) *Sleep, Wakefulness, and Circadian Rhythm*. NATO Advisory Group for Aerospace Research and Development, Paris, 1979: 8.1–8.17.

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

Johnson, L.C. and Naitoh, P. *The Operational Consequences of*

Sleep Deprivation and Sleep Deficit (NATO AGARDograph No.193). NATO Advisory Group for Aerospace Research and Development, Paris, 1974.

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

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

Lichtenstein, S. and Fischhoff, B. Do those who know more also know more about how much they know? The calibration of probability judgments. *Org. Beh. & Hum. Perf.*, 1977, 20: 159–183.

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

Link, S.W. *The Wave Theory of Difference and Similarity*. L. Erlbaum and Associates, Hillsdale, 1992.

Morris, C.C. Retrieval processes underlying confidence in comprehension judgments. *J. Exp. Psychol.: Learn., Mem. & Cognit.*, 1990, 16: 223–232.

Murphy, A.H. A new vector partition of the probability score. *J. Appl. Meteorol.*, 1973, 12: 595–600.

Myers, J.L. and Well, A.D. *Research Design and Statistical Analysis*. Harper Collins, New York, 1991.

Naitoh, P. Sleep deprivation in human subjects: A reappraisal. *Waking & Sleeping*, 1976, 1: 53–60.

Naitoh, P. and Angus, R.G. Napping and human functioning during prolonged work. In: D.Dinges and R.Broughton (Eds) *Sleep and Alertness: Chronobiological, behavioral, and medical aspects of napping*. Raven Press, New York, 1989: 221–246.

Naitoh, P. and Townsend, R.E. The role of sleep deprivation in human factors. *Human Fact.*, 1970, 12: 575–585.

Nelson, T.O. A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychol. Bull.*, 1984, 95: 109–133.

Norton, R. The effects of acute sleep deprivation on selective attention. *Br. J. Psychol.*, 1970, 61: 157–161.

Paese, P.W. and Sniezek, J.A. Influences on the appropriateness of confidence in judgment: Practice, effort, information, and decision-making. *Org. Behav. & Hum. Dec. Proc.*, 1991, 48: 100–130.

Peirce, C.S. and Jastrow, J. On small differences of sensation. *Mem. Nat. Acad. Sci.*, 1884, 3: 75–83.

Perfect, T.J., Watson, E.L. and Wagstaff, G.F. Accuracy of confidence ratings associated with general knowledge and eyewitness memory. *J. Applied Psychol.*, 1993, 78:144–147.

Petrusic, W.M. Semantic congruity effects and theories of the comparison process. *J. Exp. Psychol.: Hum. Perc. & Perf.*, 1992, 18: 962–986.

Petrusic, W.M. and Baranski, J.V. Semantic congruity effects in perceptual comparisons. *Percept. & Psychophys.*, 1989a, 45: 439–452.

Petrusic, W.M. and Baranski, J.V. Context, context shifts, and semantic congruity effects in comparative judgments. In: D. Vickers and P. Smith (Eds) *Human Information Processing: measures, mechanisms, and models*. Elsevier, North Holland, 1989b: 231–251.

Phillips, L.D. *Bayesian Statistics for Social Scientists*. Nelson, London, 1973.

Pigeau, R.A., Angus, R.G. and Heslegrave, R.J. Electrophysiological measures of mental fatigue and declining performance resulting from sleep loss. In: *Proc. 29th Ann. Conf. of the Military Testing Ass.*, 1987a: 584–589.

Pigeau, R.A., Heslegrave, R.J. and Angus, R.G. (1987b). Psychophysiological measures of drowsiness as estimators of mental fatigue and performance degradation during sleep deprivation. In: *Electric and Magnetic Activity of the Central Nervous System: research and clinical applications in aerospace medicine*. NATO Advisory Group for Aerospace Research and Development, Paris, 1987b: 21.1–21.16.

Sanders, A.F. and Reitsma, W.D. Lack of sleep and covert orienting of attention. *Acta Psychol.*, 1982, 52: 137–145.

Savage. L.J. *The Foundations of Statistics*. Wiley, New York, 1954.

Schneider, S.L. and Laurion, S.K. Do we know what we've learned from listening to the news? *Mem. & Cognit.*, 1993, 21: 198–209.

Smith, V.L., Kassin, S. and Ellsworth, P.E. Eyewitness accuracy and confidence: Within-versus between-subjects correlations. *J. Appl. Psychol.*, 1989, 74: 356–359.

Stock, W.A., Kulhavy, R.W. and Pridemore, D.R. Responding to feedback after multiple-choice answers: The influence of response confidence. *Q. J. Exp. Psychol.*, 1992, 45: 649–667.

Vickers, D. *Decision Processes in Visual Perception*. Academic Press, New York, 1979.

Vickers, D. and Packer, J. Effects of alternating set for speed versus accuracy on response time, accuracy, and confidence in a unidimensional discrimination task. *Acta Psychol.*, 1982, 50: 179–197.

Vickers, D., Smith, P, Burt, J. and Brown, M. Experimental paradigms emphasising state or process limitations: II. Effects on confidence. *Acta Psychol.*, 1985, 59: 163–193.

Volkmann, J. The relation of the time of judgment to the certainty of judgment. *Psychol. Bull.*, 1934, 31: 672–673.

Wallsten, T.S. and Budescu, D.V. Encoding subjective probabilities: A psychological and psychometric review. *Management Sci.*, 1983, 29: 151–173.

Weaver, C.A. Constraining factors in calibration of comprehension. *J. Exp. Psychol.: Learn., Mem. & Cognit.*, 1990, 16: 214–222.

Wells, G.L., Lindsay, R.C.L. and Ferguson, T.J. Accuracy, confidence and juror perceptions in eyewitness identification. *J. Applied Psychol.*, 1979, 64: 440–448.

Wilkinson, R.T. Sleep deprivation. In: O. Edholm and A.Bacharach (Eds) *The Physiology of Human Survival*. Academic Press, New York, 1965: 399–430.

Wilkinson, R.T. Some factors influencing the effect of environmental stresses upon performance. *Psychol. Bull.*, 1969, 72: 260–272.

Wimmer, F., Hoffmann, R.F., Bonato, R.A. and Moffitt, A.R. The effects of sleep deprivation on divergent thinking and attention processes. *J. Sleep. Res.*, 1992, 1: 223–230.

Yaniv, I., Yates, J.F. and Smith, J.E.K. Measures of discrimination skill in probabilistic judgment. *Psychol. Bull.*, 1991, 110: 611–617.

Yates, J.F. External correspondence: Decompositions of the mean probability score. *Org. Behav. & Hum. Dec. Proc.*, 1982, 30: 132–156.

Yates, J.F. *Judgment and Decision Making*. Prentice Hall, Englewood Cliffs, 1990.

#389685

NO. OF COPIES NOMBRE DE COPIES	COPY NO. COPIE N°	INFORMATION SCIENTIST'S INITIALS INITIALES DE L'AGENT D'INFORMATION SCIENTIFIQUE
1	1	JC
AQUISITION ROUTE FOURNI PAR ▶	DCIEM	
DATE ▶	20 MARCH 1996	
DSIS ACCESSION NO. NUMÉRO DSIS ▶		

DND 1158 (6-87)

National Defence / Défense nationale

PLEASE RETURN THIS DOCUMENT TO THE FOLLOWING ADDRESS:

DIRECTOR
SCIENTIFIC INFORMATION SERVICES
NATIONAL DEFENCE
HEADQUARTERS
OTTAWA, ONT. - CANADA K1A 0K2

PRIÈRE DE RETOURNER CE DOCUMENT À L'ADRESSE SUIVANTE:

DIRECTEUR
SERVICES D'INFORMATION SCIENTIFIQUES
QUARTIER GÉNÉRAL
DE LA DÉFENSE NATIONALE
OTTAWA, ONT. - CANADA K1A 0K2