A Meta-Analysis of the Impact of Short-Term Sleep Deprivation on Cognitive Variables

Julian Lim and **David F. Dinges**
Division of Sleep and Chronobiology, Department of Psychiatry, University of Pennsylvania School of Medicine

Abstract

A substantial amount of research has been conducted in an effort to understand the impact of short-term (<48 hr) total sleep deprivation (SD) on outcomes in various cognitive domains. Despite this wealth of information, there has been disagreement on how these data should be interpreted, arising in part because the relative magnitude of effect sizes in these domains is not known. To address this question, we conducted a meta-analysis to discover the effects of short-term SD on both speed and accuracy measures in 6 cognitive categories: simple attention, complex attention, working memory, processing speed, short-term memory, and reasoning. Seventy articles containing 147 cognitive tests were found that met inclusion criteria for this study. Effect sizes ranged from small and nonsignificant (reasoning accuracy: $\bar{g} = -0.125$, 95% CI [−0.27, 0.02]) to large (lapses in simple attention: $\bar{g} = -0.776$, 95% CI [−0.96, −0.60], $p < .001$). Across cognitive domains, significant differences were observed for both speed and accuracy; however, there were no differences between speed and accuracy measures within each cognitive domain. Of several moderators tested, only time awake was a significant predictor of between-studies variability, and only for accuracy measures, suggesting that heterogeneity in test characteristics may account for a significant amount of the remaining between-studies variance. The theoretical implications of these findings for the study of SD and cognition are discussed.

Keywords

sleep deprivation; meta-analysis; attention; working memory; processing speed

Since the seminal experiments on human sleep deprivation (SD) in the late 19th century (Patrick & Gilbert, 1896), there have been frequent attempts to better understand the effects of SD on human physiology and performance. Within this body of research, substantial resources have been invested in interrogating the effects of total SD (complete sleep loss over an extended period) on cognition and cognitive performance. Though less common than long-term chronic sleep loss, in which sleep time is partially reduced over days or weeks, cognitive failures associated with total SD are nevertheless of great interest and importance, as their real-world consequences are often catastrophic (Dinges, 1995; Mitler et al., 1988).

Within the literature, much attention has centered on the effects of SD on basic attentional processes and complex, real-world tasks. There has been relatively less interest in other cognitive domains on which SD has a known deleterious effect, such as short-term memory,

Correspondence concerning this article should be addressed to Julian Lim, Unit for Experimental Psychiatry, Department of Psychiatry, University of Pennsylvania School of Medicine, 1020 Blockley Hall, 423 Guardian Drive, Philadelphia, PA 19104. julianzl@sas.upenn.edu.

working memory, processing speed, and tasks of selective attention. Although there is general agreement that SD does exert an effect across most of these cognitive domains (for a review, see Durmer & Dinges, 2005), much less is known about the degree of these negative consequences. Moreover, various studies have found little to no effect of short-term SD on tests within some of these cognitive domains (e.g., Binks, Waters, & Hurry, 1999; Quigley, Green, Morgan, Idzikowski, & Kin, 2000), which has led to some disagreement about how to fit these data into a coherent theory.

As a means of performing this synthesis, several comprehensive qualitative reviews on the effects of SD on cognitive function have appeared over the past several years. By and large, these reviews have concluded that SD has effects on a broad suite of cognitive domains, but they hesitate to make claims about the relative magnitude of these effects. Moreover, there exist three general schools of thought as to the theoretical framework through which the available data should be interpreted. We summarize these viewpoints below, noting that these theories are not mutually exclusive and certainly amenable to integration.

The Controlled Attention Hypothesis

Many of the early studies on SD and cognition cite novelty and motivation as critical variables in determining performance under adverse conditions (Wilkinson, 1961; Williams, Lubin, & Goodnow, 1959). These suggestions were made subsequent to the initially paradoxical observations that many highly demanding cognitive tests are unaffected by short bouts of total SD. For example, performance on Baddeley's Logical Reasoning Test is consistently found to be stable, even as sleepiness and impairment in other cognitive domains appear (Magill et al., 2003; A. P. Smith & Maben, 1993). These negative findings have prompted the creation of theories such as the "controlled attention" model of Pilcher, Band, Odle-Dusseau, and Muth (2007). In this model, the authors highlighted the importance of "bottom-up" task characteristics, arguing that tasks that are monotonous or intrinsically less engaging are more severely affected by SD, due to the fact that greater top-down control is needed to sustain optimal performance on these tests. The authors suggested that tasks be classified on the basis of whether they encourage attentive behavior and hypothesized that tasks that are high on this dimension are affected the least by SD.

The Neuropsychological Hypothesis

Several reviewers have suggested that SD has domain-specific effects on cognition, with particular focus on tasks mediated by prefrontal cortex (PFC) function. Jones and Harrison (2001) and Harrison and Horne (2000) both reviewed the literature on the impact of SD on PFC-oriented tasks and concluded that these tests provide incremental validity in assessing impairment beyond the consideration of vigilance or sustained attention alone. For example, Harrison, Horne, and Rothwell (2000) gave young adults a neuropsychological battery following 36 hr of total SD and found specific impairments on PFC-oriented tests (of temporal memory, verbal fluency, and response inhibition) but not on a test of recognition memory. The authors noted that the impairments seen were similar to those displayed by healthy, middle-aged (55–64 years) participants, with diminution of PFC function being a known consequence of normal aging. More recently, neuroimaging data have lent further support to this claim; for instance, studies using functional magnetic resonance imaging (Chee & Choo, 2004; Drummond et al., 2000) have demonstrated hypoactivation in regions of the lateral and medial PFC to a variety of tasks following SD, thus localizing the putative neural basis for the observed behavioral changes.

Proponents of this view interpret these findings as evidence that the impairment seen in many complex cognitive tasks is not merely driven by the failure of more basic cognitive skills; that is, PFC-oriented tasks are vulnerable to specific failures that are above and

beyond those expected to be caused by low arousal and sleepiness (Harrison et al., 2000). Conceptually, this can be thought of as a neuropsychological model; that is, SD produces a reversible functional lesion in the PFC that is detectable by tests sensitive to these deficits in brain-injured patients. This model provides some explanatory power in resolving the mixed results in the literature that researchers had tried to account for with moderators such as task type, task length, novelty, and motivation.

The Vigilance Hypothesis

Finally, other reviewers have singled out arousal and vigilance as general factors that explain much of the variance in cognitive deficits following sleep loss. Durmer and Dinges (2005) stated that "cognitive tasks vary considerably in their sensitivity to sleep loss" (p. 120) but remarked that reaction time measures of tasks of attention and vigilance are the predominant instruments used to assay vulnerability to SD. Lim and Dinges (2008) also spotlighted vigilant attention as a cognitive process that is consistently and robustly affected by total SD. Finally, Balkin, Rupp, Picchioni, and Wesensten (2008) made the stronger assertion that "the array [of activities affected by sleep loss] is so extensive that it is reasonable to posit that sleep loss exerts a nonspecific effect on cognitive performance" (p. 654).

There is strong experimental evidence for these assertions. Tests of sustained attention (e.g., the Psychomotor Vigilance Test) are not only reliable but also highly valid in predicting real-world performance and assessing the level of impairment faced by an individual under conditions of fatigue (Dinges et al., 1997; Lim & Dinges, 2008). The Psychomotor Vigilance Test is also sensitive in tracking both circadian and homeostatic modulations in sustained attention and arousal over the course of several days without sleep (Doran, Van Dongen, & Dinges, 2001). Finally, models of attention often stress that vigilance and sustained attention are fundamentally important to many higher aspects of cognition and that these higher processes will necessarily decline if a subject is not able to sustain a sufficient level of vigilance while performing a task (Sturm et al., 1999; Sturm & Willmes, 2001).

The three models discussed are above not mutually incompatible. One could argue that the controlled attention hypothesis and the vigilance hypothesis merely take different perspectives in explaining the same set of phenomena and that the neuropsychological hypothesis, though consistent with both of these models, accounts for effects above and beyond what may be expected from either. As a result, certain theorists have proposed a more integrative approach in interpreting the available data. For instance, Boonstra, Stins, Daffertshofer, and Beek (2007) suggested that impairment in the PFC following a period of SD may underlie changes in both executive functioning and attention, stressing the role of the PFC in the interaction between top-down and bottom-up processes.

If we believe that there is some predictive power in all of these models, a new and more pragmatic question arises: To what degree are different cognitive functions impaired? Without a standardized basis of comparison, there is no way to assess the relative importance of each of these theoretical approaches. Knowledge of the effect sizes associated with each of these impairments may be of use in determining the targets for intervention in real-life situations so as to minimize the deleterious effects of SD on workers in the field.

The meta-analysis (M. L. Smith & Glass, 1977) is an increasingly popular method of synthesizing data from the primary literature and is a useful tool in addressing the question posed above. This method entails a systematic search for all articles related to a topic that meet a preordained set of inclusion criteria, calculating the effect sizes in all studies that meet these criteria and accumulating these effect sizes by weighting them on the basis of their sample sizes. This final step uses an estimate of sampling variance to give greater

weight to studies with larger samples, thus providing a more unbiased estimate of the true effect size of a given manipulation.

Prior Meta-Analyses

To our knowledge, three meta-analyses have been conducted to date on the effects of SD on performance. Koslowsky and Babkoff (1992) summarized 27 studies and reported that total SD showed greater correlations with performance as the duration of deprivation increased, and that speed or latency variables were generally affected more than accuracy measures, arguing that these data support the lapse hypothesis of Williams et al. (1959). Pilcher and Huffcutt (1996) analyzed 19 studies for the effects of SD on cognitive task performance, motor task performance, and mood and concluded that "simple" tasks were affected more than "complex" tasks after short periods of SD but that the reverse was true for long periods (>45 hr) of SD. The same pattern was seen with task duration, with short tasks being more adversely affected after short periods of SD, and the reverse true of long tasks. These authors found that partial SD (in which a restricted amount of sleep is allowed every night) had a more pronounced effect on cognition overall than total SD. Most recently, Philibert (2005) conducted a meta-analysis to assess the effects of SD on cognitive and clinical performance in physicians and nonphysicians. Overall, the effect sizes for changes in cognitive function were −0.564 (95% CI [−0.406, −0.722]), with the most pronounced effects on tests of vigilance and sustained attention ($d = -1.33$, 95% CI [−1.124, −1.536]). This final analysis was the most comprehensive and methodologically sophisticated of the three, with 60 studies and 5,295 individual effect indices included.

The studies described above have a number of weaknesses that remain to be addressed. First, two of these meta-analyses (excepting Koslowsky and Babkoff, 1992) aggregated performance variables measuring accuracy and speed into a single category when summarizing effect sizes. There is little empirical evidence that speed and accuracy are uniformly affected by SD, and aggregating the two outcome types may result in the loss of interesting information. Second, the cognitive categories in these previous analyses were relatively coarse; for example, no distinction was made in the Philibert (2005) analysis between working memory and short-term memory paradigms. Finally, none of the previous analyses performed attempted to control for differences in study quality or took into account the interindividual differences present in cognitive responses to total SD.

Our purpose in conducting the current meta-analysis was thus twofold: (a) to investigate the relative magnitude of the effects of SD on different cognitive domains and (b) to explore whether the effects on accuracy and reaction time measures were different in any of these domains. The overarching motivation for this analysis was to uncover evidence that may inform our understanding of the effects of short-term acute SD on cognition and thus aid in assessing the relative importance of current theoretical models.

Method

Study Selection

Our primary collection of literature was gathered by searching online electronic databases for articles relevant to our topic of interest through December 2008. The four databases used were PsycINFO, Medline, Web of Science, and EBSCO MegaFile. In each of these databases, we conducted a search using a combination of the following terms: *sleep deprivation* or *sleep restriction* and *cognition*, *attention*, *memory*, *performance*, *vigilance*, and *executive function* (12 combinations in all). This search yielded 4,276 hits in total. We next scanned the abstracts of these articles to determine their suitability for inclusion in the analysis. In total, 176 of the articles were empirical studies that employed SD as a primary

independent variable and used at least one cognitive measure as a dependent outcome. We subsequently obtained the full text of these articles to determine whether they met full inclusion criteria. These criteria were as follows:

1. Participants in the study must all have been healthy adults aged 18 years and older.

2. The study must have included as its primary manipulation a specified length of total SD not exceeding 48 hr.

3. The study must have included as a dependent measure at least one test of basic cognitive function, and the description of the test must have been specific enough for us to classify it as an assay of a particular cognitive domain (we elaborate on this point further below).

4. There was sufficient statistical information in the study for the calculation of effect sizes.

Because of the restrictions imposed by Criterion 3, a number of subareas within the realm of SD research necessarily had to be omitted from this analysis. A survey of the literature on SD and decision making revealed that outcome variables on these tests did not form a cluster that was homogeneous enough to warrant a quantitative synthesis. This was because many of these experiments employed complex, real-world scenarios, opting for ecologically valid designs over more controlled neuropsychological tests (for a review, see Harrison & Horne, 2000). Moreover, it is unclear how outcome variables from standardized decision-making tests (e.g., the Iowa Gambling Test) should be compared with the accuracy measures obtained from other cognitive domains. Finally, experiments on implicit and procedural memory were also excluded, as these form a separate body of literature pertaining to sleep and memory consolidation (Stickgold & Walker, 2005), the analysis of which is beyond the scope of this article.

In addition to this online literature search, we obtained data from several other sources. We conducted hand searches of the journal *Sleep* and the *Journal of Sleep Research* from 1988 to 2008. We also reviewed the reference lists of the major review articles on SD and cognitive performance that have been published over the last several years. Finally, to mitigate the "file drawer" problem (Strube & Hartmann, 1983), we contacted eight major sleep laboratories conducting research in this field to request unpublished data from experiments, as well as master's and doctoral theses. We received additional data from one of these laboratories, as well as replies from all but one of the remaining investigators informing us that they did not have suitable data for inclusion. In total, 70 articles and 147 data sets met inclusion criteria and were included in the meta-analysis (see Table 1). Among these, numerous data sets contained more than one cognitive outcome; these were coded separately, according to the recommendations of Hunter and Schmidt (2004). Altogether, 209 aggregated effect sizes and 5,104 individual effect sizes were calculated from these data sets.

Cognitive Domains

Each cognitive test was assigned a domain according to the classification scheme below.

Simple attention—Tests in simple attention involved the visual or auditory detection of a single class of stimuli, with no element of perceptual discrimination, orienting, inhibition, or any other form of executive attention, such as the Psychomotor Vigilance Test (Dinges & Powell, 1985) and other simple reaction time tests. This was the only category in which effect sizes were calculated for lapses and omissions instead of accuracy.

Complex attention—Tests in complex attention assessed all attentional processes more demanding than those in the first category (e.g., selective or executive attention) but did not have any major working memory component or require any short-term or long-term memory encoding, such as the Stroop test (Stroop, 1935), the Continuous Performance Test (Conners, 1995), and the go/no-go paradigm.

Processing speed—Tests in processing speed primarily assessed cognitive throughput or processing speed, requiring multiple repetitions of a rehearsed process within a fixed period. Examples include the Digit Symbol Substitution Test from the Wechsler Adult Intelligence Scale (Wechsler, 1997a) and tests of serial addition and subtraction.

Working memory—Tests in working memory involved the maintenance and/or manipulation of relevant information over a brief period, culminating in a decision and response, such as the Sternberg working memory task and the delayed-match-to-sample test.

Short-term memory—Tests in short-term memory involved the encoding, maintenance, and retrieval of information. The amount of information to be stored had to exceed working memory capacity, and maintenance typically occurred over a longer period. Examples include word list learning and the Logical Memory subtest of the Wechsler Memory Scales (Wechsler, 1997b). This domain was further subdivided into short-term memory recall and short-term memory recognition. Only effect sizes for accuracy measures were computed for this cognitive category. We note that individuals with a background in cognitive psychology may consider many of these as long-term memory tests due to differences in nomenclature across fields.

Reasoning and crystallized intelligence—Tests in reasoning and crystallized intelligence assessed mental processes such as problem solving, vocabulary exercises, and other forms of crystallized cognitive ability. Examples include Raven's Advanced Progressive Matrices test (Raven, Raven, & Court, 1998) and Baddeley's Grammatical Reasoning Test (Baddeley, 1968). Only effect sizes for accuracy measures were computed for this cognitive category.

Verbal fluency—We had originally intended to analyze tests of verbal fluency (e.g., the Controlled Oral Word Association Test; Spreen & Strauss, 1991) as a seventh category in this study. However, as only three articles to date contained such tests and met all inclusion criteria (Binks et al., 1999; Fluck et al., 1998; Horne, 1988), we omitted this category from our final analysis.

Coding for Study Quality

It has been recommended that studies entered into a meta-analysis be coded for study quality (Chalmers et al., 1981; Detsky, Naylor, O'Rourke, McGeer, & L'Abbé, 1992). This is especially important when the pool of studies entered into the analysis is highly heterogeneous and the designs have varying threats to internal and external validity (Detsky et al., 1992). In our survey of the literature, we observed significant discrepancies in how experiments of SD are conducted and controlled, and thus deemed that this step was appropriate for our analysis. We identified seven features that were important determinants of a study's reliability and validity, and coded each experiment so that they received a 0 or 1 score on each of these criteria:

Randomization and counterbalancing—For between-subjects studies, were subjects randomized to the sleep-deprived and control groups? For repeated-measures studies, was the study order counterbalanced to avoid the potential confound of order effects?

Adequacy of control group—Were the treatment and control groups equal in number? Were they treated similarly (e.g., in compensation and study conditions)?

Subject recruitment—Were subjects randomly recruited from the population? Was the study sample representative of the population, or did the experiment include only a particular subgroup of people (e.g., fighter pilots, only women)?

Quality of statistical analysis—Were appropriate statistical tests used to analyze the data?

Adequacy of measures used—Did the cognitive tests used appropriately capture the construct of interest? Were they well validated? Were appropriate dependent measures used in the analysis?

Adequacy of control over SD—Was the study conducted in a sleep laboratory? Were subjects monitored during their time in the study? Were their diet and activity controlled?

Adequacy of control over sleep history—Were participants screened for good sleep history or absence of sleep disorders? Was sleep history monitored in the period prior to the experiment (e.g., sleep diaries or actigraphy)?

Coding Reliability

Studies were assessed and coded by two independent raters (Julian Lim and one other rater who was trained on the criteria above). They assessed interrater reliability using intraclass correlation coefficients from a two-way mixed model with raters as fixed effects and studies as random effects. The intraclass correlation coefficient for the entire sample was .798, indicating that there was a high level of agreement between the two raters on study quality over the entire sample of articles.

Other Study Variables

In addition to coding for study quality, we recorded the following variables for use as potential moderators in the secondary analysis: length of SD and the times of test administration (which were used to calculate circadian offset).

Effect Size Calculation

The primary metric of the meta-analysis is the effect size, which is a standardized estimate of the magnitude of the treatment effect. We calculated all effect sizes (in this analysis, *d* values) using a baseline test and the test at the most extreme point of SD in the experiment (with a few exceptions, noted in Table 1). In the case where means (μ) and standard deviations or errors (σ) were reported, we calculated effect sizes using Equation 1 for between-subjects studies:

$$d = \frac{\mu_2 - \mu_1}{\sqrt{\frac{N_1 \cdot \sigma^2 + N_2 \cdot \sigma_2^2}{N_1 + N_2}}},$$

(1)

where N_1 and N_2 are the number of subjects in the control and experimental groups, respectively. In contrast, effect sizes in within-subjects or repeated-measures studies should be calculated with the standard deviation of change scores as an error term (Hunter & Schmidt, 2004); however, these are seldom reported in the literature. Instead, we estimated this term using the pre- and posttest standard deviations and correlations, as suggested by Morris and DeShon (2002). In cases where this information was not available, we reverted

to Formula 1 as our best estimate of the effect size. Where only *t* or *F* values were reported, we converted these to effect sizes following the recommendations of Rosenthal (1991), as shown in Equations 2 and 3. Where only *p* values were reported (for *t* tests), we back-converted these to *t* values using statistical tables and applied Formula 3. Once a *d* value was obtained, its mathematical sign was adjusted so that worse performance following SD was always reflected by a negative effect size.

$$d = \frac{2t}{\sqrt{df}} \tag{2}$$

$$d = \frac{2 \cdot \sqrt{F_{(1,x)}}}{\sqrt{df}} \tag{3}$$

As recommended by Hedges and Olkin (1985), we next adjusted for effect size inflation in studies with small samples by applying the correction in Equation 4. This adjustment yields the Hedges's *g*, which treats the variance of each study as an estimate rather than a constant. The difference in this correction for between- and within-subjects studies is accounted for by the differing degrees of freedom in the denominator of the equation.

$$g = d \times \left(1 - \frac{3}{4.df - 1} \right) \tag{4}$$

In order to combine the results of between- and within-subjects designs, all effect sizes need to be converted into a single metric. As the effect of interest in this study was that of SD on an individual over time, the within-subjects design is the appropriate standard of measurement for this analysis. Effect sizes from between-subjects experiments were converted into the appropriate metric by the formula suggested by Morris and DeShon (2002; see Equation 5), where ρ is the correlation between the pre- and posttest scores.

$$g_{WS} = \frac{g_{BS}}{\sqrt{2(1 - \rho)}} \tag{5}$$

As ρ is seldom reported in the literature, we estimated this value from data collected on various cognitive tests in our own laboratory. Pre- and posttest correlations from this investigation generally fell in the .4 –.6 range. To simplify this conversion, therefore, we assumed that for SD experiments, ρ = .5 (the unity case where $g_{ws} = g_{bs}$).

Calculation of Sampling Variance

We computed the within-studies variance due to sampling error (Var(*e*)) for each of the data sets using Equation 6 for within-subjects studies and Equation 7 for between-subjects studies. For within-subjects studies, we used a sampling variance term that takes into account a Treatment × Subjects interaction[1] (Hunter & Schmidt, 2004), as recent research has emphasized the large and stable intersubject variability in cognitive performance after SD (Leproult et al., 2003; Van Dongen, Baynard, Maislin, & Dinges, 2004).

---

[1]This is accounted for by *a* in Equation 6, where $a = 2(1 − \rho)/r_{YY}$ and r_{YY} is the square root of the test–retest reliability. In cases where reliability information for a particular test was not available, we first searched the literature for tests that were highly similar to the one used, then as a last resort used the average reliability from tests within the respective cognitive domain. In all cases, separate reliability coefficients were located and used for accuracy and reaction time measures.

$$\mathrm{Var}(e) = \left[\frac{N-1}{N-3}\right] \cdot \left[\frac{1}{N}\right] \cdot \left[\frac{a^2+g^2}{2}\right] \tag{6}$$

$$\mathrm{Var}(e) = \left[\frac{N-1}{N-3}\right] \cdot \left[\frac{4}{N}\right] \cdot \left[1+\frac{g^2}{8}\right] \tag{7}$$

Meta-Analysis Procedure

Separate analyses were conducted for accuracy (or lapses) and speed for the cognitive domains of simple attention, complex attention, working memory, and processing speed. Only accuracy measures were compiled for the domains of short-term memory and reasoning.

We calculated the overall average effect size for each outcome measure type and domain using Equation 8. Each effect size was weighted by the inverse of its sampling variance (w_i), thus giving relatively less weight to studies with small sample sizes. Effect sizes were also weighted by their individual mean-adjusted study quality (sq_i; i.e., quality for study i divided by the mean quality rating in its respective domain). Mean-adjusted scores are a viable method for accounting for differences in study quality[2] (Bérard & Bravo, 1998; Detsky et al., 1992) and have the advantage of not widening the confidence intervals of pooled effect sizes.

$$\bar{g} = \sum d_i w_i sq_i / \sum w_i sq_i \tag{8}$$

The weights (w_i) in Equation 8 were derived via a random-effects model, which assumes that the "true" effect size of each study is not identical and that samples were drawn from populations that differ in meaningful ways. This was clearly the case in our analysis; as a most basic example, the length of SD varied from 24 to 48 hr between studies, and it is known that the magnitude of performance deficits grows with escalating sleep pressure (Doran et al., 2001).

Finally, by calculating the variance components associated with between-studies and within-studies variation, we were able to obtain two measures of dispersion for each set of studies, the Q statistic, which reflects the total amount of variance in the meta-analysis, and the I^2 statistic, which ranges from 0 to 100 and is an index of the proportion of variance in the sample attributable to between-studies differences (Higgins, Thompson, Deeks, & Altman, 2003).

Results

A complete list of studies and individual effect sizes is presented in Table 1. The total sample size for the analysis was 1,533, with an average of 21.3 ($SD = 11.1$) subjects in each study. The average study quality for the complete sample ranged from 3 to 7 ($M = 5.21$, $SD = 1.18$).

---

[2]However, note that there is no gold standard as yet of incorporating study-quality information into pooled effect sizes. For the purposes of comparison, the supplemental materials table reports pooled effect sizes for each cognitive domain with and without these study-quality weights.

Aggregate Effect Sizes

Average effect sizes for each cognitive domain and outcome are presented in Table 2. A total SD period of 24 – 48 hr had a significant effect in reducing performance for outcomes in all cognitive domains, except for accuracy measures in tasks of processing speed (p = .06), as well as accuracy measures on tests of reasoning and crystallized intelligence (p = .08). As there were relatively few studies in each of these categories, however, it is possible that the analysis had insufficient power to detect a significant effect for these outcomes.

As anticipated, the largest effects of 24 – 48 hr of SD were on tests of vigilance, or simple attention. Effect sizes within this domain were −0.762 (lapses) and −0.732 (reaction times), which represent moderate to large effects. These effects are comparable to those reported by Philibert (2005), who found effect sizes of −1.142 and −0.553 for vigilance tests conducted after 24 –30 and 30 –54 hr of SD, respectively. Effect sizes for complex attention and working memory fell in the moderate range, and tests of processing speed showed on average small but significant effects.

We performed analyses of variance on the aggregate effect sizes to test two sets of null hypotheses: first, that average effect sizes are identical across cognitive domains (with separate analyses conducted for speed and accuracy), and second, that average effect sizes for speed and accuracy are identical within each cognitive domain. As two of the cognitive domains (short-term memory and reasoning) contained only one outcome measure, we did not enter all information into a two-way analysis of variance. Tables 3 and 4 summarize the results of this analysis. We found a significant difference in effect sizes across cognitive domains for measures of both speed, $Q(3) = 24.5$, $p < .001$, and accuracy, $Q(5) = 36.8$, $p < .001$; however, there were no differences between speed and accuracy measures within each cognitive domain, even prior to correction for multiple comparisons.

The I^2 value is an index of the proportion of variance within each analysis that is due to between-studies differences; Higgins et al. (2003) suggested that values of 25, 50, and 75 may be used as benchmarks of low, moderate, and high heterogeneity, respectively. I^2 values in each analysis ranged from small (reasoning: 5.38) to moderate to high (working memory reaction time: 66.28; see Table 2), indicating that moderator analyses was appropriate in most of these cognitive domains. Indeed, given that the number of hours of SD in these studies ranged from 24 to 48, and that several types of cognitive tests made up these collective indices, it would have been surprising to discover low I^2 values in this first-pass analysis. As a way to visualize this dispersion, Figures 1 and 2 are displays of forest plots of the accuracy and reaction time measures for a sample domain (simple attention).

Moderator Analyses

We coded three study variables to test their impact as moderators of the effect of SD. Circadian time was estimated by plotting the time of test administration as a sinusoidal function with a 24-hr period and a performance nadir at 0600 hr, with peak amplitude arbitrarily defined as 1. Circadian offset was computed by subtracting the time of test administration for sleep-deprived subjects from time of test administration of the control group. Homeostatic sleep pressure was estimated as the elapsed time between sleep offset and time of test administration. In cases where any of this information was not explicitly reported, or the testing time occurred over a range greater than 2 hr, we did not code these variables, and the effect size was excluded from the moderator analysis.

As there were insufficient cases to conduct separate metaregressions within each cognitive category, we combined all results for accuracy and reaction time effects, and conducted stepwise multiple regression within these two larger data sets, using the average effect size found for each cognitive domain as a covariate. For accuracy measures, the omnibus effect

for the model was significant, $R^2 = .176$, $F(2, 97) = 10.39$, $p < .001$, but only homeostatic sleep pressure was a significant predictor of study effect size, $b = -0.22$, $t(98) = -2.43$, $p = .02$. In contrast, the overall model for reaction time measures was not significant, indicating that none of the coded variables were a significant predictor of heterogeneity in this sample.

Discussion

The results from our meta-analysis support the conclusions of previous reviews that short-term total SD has a significant deleterious effect across most cognitive domains. Our current study represents an advance over previous meta-analyses in several important respects. First, we were able to take into account the known Treatment × Subject interaction in experiments of SD (Van Dongen et al., 2004), thus improving the estimation of the sampling variance for each study. Second, we weighted each effect size on the basis of study quality, thus giving less influence to studies that may have been less well conducted. Third, we had more stringent inclusion criteria than Philibert (2005), which increased the homogeneity of our sample. Finally, and most important, we classified behavioral tests into finer grained cognitive domains than previous meta-analyses, further increasing the similarity of studies within each subsample.

Overall, average effect sizes appear to fall along a continuum, with tasks of greater complexity affected relatively less after total SD. The relative magnitude of effect sizes across cognitive domains was similar to those seen in the meta-analysis of Philibert (2005), although the absolute size of these effects was smaller across all categories. This is likely due to two reasons: We excluded all studies with a period of total SD greater than 48 hr, and we did not disattenuate effect sizes based on test–retest reliability of dependent measures.

The difference in the average effect size among the six cognitive domains was statistically significant and ranged from −0.125 to −0.762. As anticipated, the combined effect size for simple attention and vigilance tasks was the largest among all the categories studied. This finding is consistent with the notion that vigilance is the fundamental process affected by SD (Lim & Dinges, 2008) and the deficit for which compensation is least available. In contrast, average effect sizes for complex attention and working memory tests fell into the moderate range. Although this pattern of results has been observed in the literature, this is, to our knowledge, the first time that this difference has been systematically investigated in a large body of studies.

Several points of interest arise on inspection of the group effect sizes of the complex cognitive tasks (all categories other than simple attention). First, we note that task performance in the complex attention category is relatively spared when compared with simple attention. These data are compelling, as many of the complex attention tests differ from the simple attention tests in only a single cognitive process (e.g., two-choice reaction time vs. simple reaction time). This finding suggests that for tests of orienting or executive attention, performance is relatively preserved after SD either because of the greater salience of the bottom-up feed (and thus the reduced need for internally motivated top-down control) or because of the recruitment of additional mental operations. However, we also observe that complexity alone is an inadequate construct with which to identify tasks that may not be as affected by SD, as there were still substantial effect size differences among complex tasks in different domains. The nuances of these behavioral effects, as well as their neural correlates, should continue to be an interesting and fruitful area of study.

We failed to find significant effects in two of the categories tested. First, there was no effect of SD on accuracy measures in tests of reasoning and crystallized intelligence. Crystallized abilities (e.g., the retrieval of domain-specific knowledge) are thought to be highly stable

over a range of cognitive states, and are even of use in assessing premorbid functioning following neurological insult or the onset of dementia (O'Carroll & Gilleard, 1986; Watt & O'Carroll, 1999). It is unsurprising, therefore, that outcomes on these tests are relatively unaffected by short-term SD.

Second, the average effect size of the change in accuracy measures for tests of processing speed failed (but only barely) to reach statistical significance. There are at least two potential explanations for this finding. Nearly all the tasks in the processing speed category were self-paced, as opposed to work paced, and several authors have commented on the differences between these two classes of tests. Williams et al. (1959) noted that a bias toward accurate responding is commonly found in complex, self-paced assignments, a conclusion reiterated by more recent researchers who have found speed but not accuracy effects on these tasks (e.g., De Gennaro, Ferrara, Curcio, & Bertini, 2001). Koslowsky and Babkoff (1992) also found a similar effect of work- versus self-paced tasks in their meta-analysis, although this increased effect size was seen only in studies with more than 48 hr of SD. A less common explanation of the relatively preserved accuracy on processing speed tasks relates to the nature of the operations being performed in them. These operations usually involve high levels of automaticity (e.g., decoding symbols in the Digit Symbol Substitution Test), and the fidelity of such overlearned skills is probably protected even during periods of fatigue, leading to the relatively small increase in the number of errors made.

An important feature of the current meta-analysis was the separate aggregation of accuracy and reaction time measures. Although there is some evidence that lapsing and lapse duration after SD are correlated in a test of simple reaction time (Lim & Dinges, 2008), there is no a priori reason to assume that this relationship should hold across all cognitive domains. This point is not intuitive and warrants further discussion. Figure 3 illustrates the curve representing the speed–accuracy trade-off in a typical cognitive test, as well as the downward shift of this curve following a period of SD. The unexplored factor in this relationship is whether SD also biases subjects toward faster or slower responding, as represented by a shift along the lower curve. For instance, increases in the number of commission errors or false alarms on simple reaction time tests after SD have been attributed to increased disinhibition (Dorrian et al., 2005), which can be thought of as a bias toward faster (and less accurate) responding.

As it turns out, the results of our analysis show remarkable agreement between accuracy and reaction time measures in each cognitive category: Overall, there was no significant effect when comparing accuracy and reaction time across the sample. This finding suggests that, on average, SD does not bias subjects toward either faster or more accurate responding, although this claim cannot be made of any individual cognitive test.

Moderator Analysis

Of the three moderator variables studied, only hours awake (homeostatic sleep drive or sleep pressure) was a significant moderator of the effect of SD, and only for accuracy, not reaction time variables. Because of the nature of the coding in this study, we expected homeostatic sleep pressure to be a stronger predictor than circadian time or circadian offset, as there is considerable variability in endogenous circadian phase across individuals (Horne & Ostberg, 1976). Nevertheless, the results obtained in this analysis were surprising, as both circadian factors and homeostatic sleep drive are known to modulate cognitive performance (Mallis, Mejdal, Nguyen, & Dinges, 2004; Van Dongen & Dinges, 2005).

A likely explanation for this negative result is that much of the observed heterogeneity is due to the variety of cognitive tests in each sample. If this assertion is correct, it implies that the amount of impairment on tests that putatively assess the same cognitive domain may still

differ considerably following SD. In other words, the validity of these tests in assessing the cognitive process may not be as high after SD. For example, total SD is known to exacerbate the time-on-task effect (Doran et al., 2001), suggesting that test length may be a confounding variable across tests of many cognitive processes. To obtain an objective standard of impairment, therefore, it may be necessary to establish norms on several of the most commonly used tests in each domain.

Although it would have been interesting to test the moderating effect of self-paced and work-paced paradigms in this analysis, these variables were highly confounded with cognitive domain (i.e., within each category, most or all tests tended to be either self-paced or work paced). From the data obtained in the main effects, however, we can infer that the differential effects of self-paced versus work paced on accuracy and reaction time measures are unlikely to be significant as suggested in previous meta-analyses. Instead, it is possible that these effects are present only under certain conditions (e.g., extremely long periods of SD or for particular subsets of tests).

Theoretical Implications

As stated in the introduction, the chief objective of this meta-analysis was not to rule out any particular theoretical model but to direct attention to which of these models may have the greatest importance in explaining the real-world consequences of total SD. Although total SD does produce statistically significant differences in most cognitive domains, the largest effects are seen in tests of simple, sustained attention. This form of attention is critical in many industries involving sustained operations, during which a worker's primary task may involve long, monotonous periods of low-level monitoring and situational awareness. Moreover, relatively brief failures of vigilance may potentially lead to disastrous consequences. For example, lapses in sustained attention are the direct cause of SD-related motor vehicle accidents (Dinges, Mallis, Maislin, & Powell, 1998), in which an eyelid closure of 4 s is a sufficient amount of time for a driver to completely veer off a highway. We argue, therefore, that this cognitive module is of the greatest practical concern in combating SD-related problems in real-world situations.

A striking feature of this deficit in sustained attention is how rapidly large changes emerge. Although our analysis was restricted to subjects who had gone a single night without sleep, effect sizes were still large for both speed and accuracy measures on simple attention tasks. These findings support the data showing that deficits in sustained attention often presage the other observable cognitive effects of SD and may have considerable utility as an early warning system for imminent cognitive failure. This cognitive component should therefore be one of the primary targets of assessment for work fitness and a basis for decisions on whether subsequent countermeasures should be applied.

On the next rung of the hierarchy, we note that tests of working memory and other tests of executive attention are also robustly affected by one night of SD. Considerable research has been conducted over the past several decades to assess the effects of SD on decision making and its component subprocesses (e.g., response inhibition, updating strategies, assessing risk; Harrison & Horne, 2000), and our data suggest that further investigation into these problems is a worthwhile endeavor. Indeed, neuroimaging data on these tasks are affording us new insights into the neural processes underlying the observable behavioral changes (for a review, see Chee & Chuah, 2008) and suggesting possible neuropharmacological mechanisms through which we may intervene to ameliorate these problems in individuals who are most vulnerable to sleep loss (Chuah & Chee, 2008).

Finally, although tests of processing speed and cognitive throughput such as the Digit Symbol Substitution Test are commonly used in SD paradigms, the results of this analysis

demonstrate that their effects are relatively small compared with those of other tests. Indeed, studies of partial SD have demonstrated little or no effect on cognitive throughput tasks (Casement, Broussard, Mullington, & Press, 2006; Dinges et al., 1997). The implication of this finding is that changes in processing speed may be theoretically interesting but not of great practical significance in explaining and predicting real-world cognitive failures (Monk, 2007).

Limitations

This analysis contains a small number of limitations that may have affected the validity of the conclusions drawn. As we were able to obtain only a small amount of unpublished data, it is possible that there was a bias in the analysis toward effect sizes that reached statistical significance. Nevertheless, we received a 100% response rate from laboratories surveyed, and all but one of these investigators denied possessing any unpublished data that met our inclusion criteria. We are, therefore, relatively confident that the study was not greatly affected by publication bias.

Although every effort was made in this analysis to classify studies into appropriate and meaningful categories, it is clear that with the possible exception of simple attention, pure assays of most of the cognitive domains we have identified do not exist. Moreover, there remained numerous dissimilarities among the forms and characteristics of the tests within each category (e.g., task length, task demands), particularly within the category of complex attention. As discussed, this is the most likely reason why heterogeneity was in the moderate range for almost all categories studied. Despite these drawbacks, we propose that our taxonomy is a useful heuristic for several reasons. First, significant between-categories differences were found in the meta-analysis, suggesting that we have captured meaningful constructs with the classification we employed. Second, we have stayed faithful to categories that are well defined in the neuropsychological literature. In many cases, focal deficits on these tests have been observed in patients with specific pathologies or injuries (e.g., working memory in attention-deficit/hyperactivity disorder patients; Barkley, 1997). Finally, several of the domains studied here have relatively high external validity. For instance, the challenge in simple attention tasks is similar to the real-world demands on air traffic controllers, and tasks such as the Psychomotor Vigilance Test have been shown to correlate highly with other indicators of dangerous, drowsy driving (Dinges et al., 1998; Price et al., 2003).

We were not able to study a number of moderator effects that may be important predictors of the residual intradomain heterogeneity. Task duration is likely to be one of these factors, with longer tasks associated with greater effect sizes due to the presence of the time-on-task effect. We were unable to code this moderator chiefly because many articles did not report task length and because of the variability in time to completion for all tasks that were self-paced. As we have already mentioned, the difference between self-paced and work-paced tests was highly confounded with cognitive domain, making it unfeasible to test this as a moderator. Additionally, variables such as novelty and motivation (Jones & Harrison, 2001), though potentially important in affecting test outcomes, are not easily quantified.

Finally, a substantial number of studies entered into this meta-analysis reported only accuracy or reaction time as a dependent variable in their final published work. As a result, we could not conduct paired comparisons of these measures to assess their reliability. We encourage authors publishing in this field in the future to consider reporting both accuracy and reaction time measures where appropriate so that their relationship after SD can be better explored. We also suggest that, wherever possible, data from individual test bouts and not just omnibus *F* values for a series of bouts be reported, so as to enable the inclusion of more studies in future quantitative syntheses.

Conclusions

The results of this analysis have revealed the pattern of effects across cognitive domains and outcomes after a period of short-term total SD. Overall, there was a significant difference among cognitive domains, but not between speed and accuracy, suggesting that SD has differential effects on different cognitive processes but does not bias subjects toward either faster or more accurate responding in any of these domains. As some of the known key moderators of this effect did not explain the remaining between-studies variance, we infer that that the remaining heterogeneity is due to intertest differences and that test characteristics can influence the level of performance in the sleep-deprived state even when they are ostensibly assessing the same cognitive domain.

Finally, our results indicate that simple attention is the cognitive domain most strongly affected by short-term SD. Although decrements in other cognitive modules such as decision-making and memory processes no doubt contribute to real-world errors and accidents, the results of this analysis argue that deficits in sustained attention may represent the most parsimonious explanation for these occurrences. Thus, in light of these and other data, we believe that countermeasures targeting this cognitive module may be the most efficient means of accident prevention in industries where SD poses a significant safety risk.

Acknowledgments

References

References marked with an asterisk indicate studies included in the meta-analysis that are discussed in the text. For a complete list, go to http://dx.doi.org/10.1037/a0018883.supp.

Baddeley AD. A three-minute reasoning test based on grammatical transformation. Psychonomic Science. 1968; 10:341–342.

Balkin TJ, Rupp T, Picchioni D, Wesensten NJ. Sleep loss and sleepiness: Current issues. Chest. 2008; 134(3):653–660. [PubMed: 18779203]

Barkley RA. Behavioral inhibition, sustained attention, and executive functions: Constructing a unifying theory of ADHD. Psychological Bulletin. 1997; 121(1):65–94. [PubMed: 9000892]

Bérard A, Bravo G. Combining studies using effect sizes and quality scores: Application to bone loss in postmenopausal women. Journal of Clinical Epidemiology. 1998; 51(10):801–807. [PubMed: 9762872]

Binks PG, Waters WF, Hurry M. Short-term total sleep deprivations does not selectively impair higher cortical functioning. Sleep. 1999; 22(3):328–334. [PubMed: 10341383]

Boonstra TW, Stins JF, Daffertshofer A, Beek PJ. Effects of sleep deprivation on neural functioning: An integrative review. Cellular and Molecular Life Sciences. 2007; 64(7–8):934–946. [PubMed: 17347797]

Casement MD, Broussard JL, Mullington JM, Press DZ. The contribution of sleep to improvements in working memory scanning speed: A study of prolonged sleep restriction. Biological Psychology. 2006; 72(2):208–212. [PubMed: 16384630]

Chalmers TC, Smith H Jr, Blackburn B, Silverman B, Schroeder B, Reitman D, Ambroz A. A method for assessing the quality of a randomized control trial. Controlled Clinical Trials. 1981; 2(1):31–49. [PubMed: 7261638]

Chee MWL, Choo WC. Functional imaging of working memory after 24 hr of total sleep deprivation. Journal of Neuroscience. 2004; 24(19):4560–4567. [PubMed: 15140927]

Chee MW, Chuah LY. Functional neuroimaging insights into how sleep and sleep deprivation affect memory and cognition. Current Opinion in Neurology. 2008; 21(4):417–423. [PubMed: 18607201]

Chuah LYM, Chee MWL. Cholinergic augmentation modulates visual task performance in sleep-deprived young adults. Journal of Neuroscience. 2008; 28(44):11369–11377. [PubMed: 18971479]

Conners, C. Conners' Continuous Performance Test. Toronto, Canada: Multi-Health Systems; 1995.

De Gennaro L, Ferrara M, Curcio G, Bertini M. Visual search performance across 40 h of continuous wakefulness: Measures of speed and accuracy and relation with oculomotor performance. Physiology & Behavior. 2001; 74(1–2):197–204. [PubMed: 11564469]

Detsky AS, Naylor CD, O'Rourke K, McGeer AJ, L'Abbé KA. Incorporating variations in the quality of individual randomized trials into meta-analysis. Journal of Clinical Epidemiology. 1992; 45(3): 255–265. [PubMed: 1569422]

Dinges DF. An overview of sleepiness and accidents. Journal of Sleep Research. 1995; 4(Suppl 2):4–14. [PubMed: 10607205]

Dinges, DF.; Mallis, MM.; Maislin, G.; Powell, JW. Final report for the US Department of Transportation, National Highway Traffic Safety Administration. Springfield, VA: National Technical Information Service; 1998. Evaluation of techniques for ocular measurement as an index of fatigue and the basis for alertness management.

Dinges DF, Pack F, Williams K, Gillen KA, Powell JW, Ott GE, Pack AI. Cumulative sleepiness, mood disturbance, and psychomotor vigilance performance decrements during a week of sleep restricted to 4–5 hours per night. Sleep. 1997; 20(4):267–277. [PubMed: 9231952]

Dinges DF, Powell JW. Microcomputer analyses of performance on a portable, simple visual RT task during sustained operations. Behavior Research Methods, Instruments, & Computers. 1985; 17:652–655.

Doran SM, Van Dongen HP, Dinges DF. Sustained attention performance during sleep deprivation: Evidence of state instability. Archives Italiennes de Biologie. 2001; 139(3):253–267. [PubMed: 11330205]

Dorrian, J.; Rogers, NL.; Dinges, DF. Psychomotor vigilance performance: Neurocognitive assay sensitive to sleep loss. In: Kushida, CA., editor. Sleep deprivation: Clinical issues, pharmacology, and sleep loss effects. New York, NY: Dekker; 2005. p. 39-70.

Drummond SPA, Brown GG, Gillin JC, Stricker JL, Wong EC, Buxton RB. Altered brain response to verbal learning following sleep deprivation. Nature. 2000; 403(6770):655–657. [PubMed: 10688201]

Durmer JS, Dinges DF. Neurocognitive consequences of sleep deprivation. Seminars in Neurology. 2005; 25(1):117–129. [PubMed: 15798944]

Fluck E, File SE, Springett J, Kopelman MD, Rees J, Orgill J. Does the sedation resulting from sleep deprivation and lorazepam cause similar cognitive deficits? Pharmacology Biochemistry and Behavior. 1998; 59(4):909–915.

Harrison Y, Horne JA. The impact of sleep deprivation on decision making: A review. Journal of Experimental Psychology: Applied. 2000; 6(3):236–249. [PubMed: 11014055]

Harrison Y, Horne JA, Rothwell A. Prefrontal neuropsychological effects of sleep deprivation in young adults: A model for healthy aging? Sleep. 2000; 23(8):1067–1073. [PubMed: 11145321]

Hedges, L.; Olkin, I. Statistical methods for meta-analysis. Orlando, FL: Academic Press; 1985.

Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ. 2003; 327(7414):557–560. [PubMed: 12958120]

Horne JA. Sleep loss and "divergent" thinking ability. Sleep. 1988; 11(6):528–536. [PubMed: 3238256]

Horne JA, Ostberg O. A self-assessment questionnaire to determine morningness–eveningness in human circadian rhythms. International Journal of Chronobiology. 1976; 4(2):97–110. [PubMed: 1027738]

Hunter, JE.; Schmidt, FL. Methods of meta-analysis: Correcting error and bias in research findings. 2. Thousand Oaks, CA: Sage; 2004.

Jones K, Harrison Y. Frontal lobe function, sleep loss and fragmented sleep. Sleep Medicine Reviews. 2001; 5(6):463–475. [PubMed: 12531154]

Koslowsky M, Babkoff H. Meta-analysis of the relationship between total sleep deprivation and performance. Chronobiology International. 1992; 9(2):132–136. [PubMed: 1533178]

Leproult R, Colecchia EF, Berardi AM, Stickgold R, Kosslyn SM, Van Cauter E. Individual differences in subjective and objective alertness during sleep deprivation are stable and unrelated. American Journal of Physiology: Regulatory, Integrative and Comparative Physiology. 2003; 284(2):R280–R290.

Lim J, Dinges DF. Sleep deprivation and vigilant attention. Annals of the New York Academy of Sciences. 2008; 1129:305–322. [PubMed: 18591490]

Magill RA, Waters WF, Bray GA, Volaufova J, Smith SR, Lieberman HR, Ryan DH. Effects of tyrosine, phentermine, caffeine D-amphetamine, and placebo on cognitive and motor performance deficits during sleep deprivation. Nutritional Neuroscience. 2003; 6(4):237–246. [PubMed: 12887140]

Mallis MM, Mejdal S, Nguyen TT, Dinges DF. Summary of the key features of seven biomathematical models of human fatigue and performance. Aviation, Space, and Environmental Medicine. 2004; 75(Suppl 3):A4–A14.

Mitler MM, Carskadon MA, Czeisler CA, Dement WC, Dinges DF, Graeber RC. Catastrophes, sleep, and public policy: Consensus report. Sleep. 1988; 11(1):100–109. [PubMed: 3283909]

Monk TH. Practical consequences of fatigue-related performance failures. Sleep. 2007; 30(11):1402–1403. [PubMed: 18041474]

Morris SB, DeShon RP. Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. Psychological Methods. 2002; 7(1):105–125. [PubMed: 11928886]

O'Carroll RE, Gilleard CJ. Estimation of premorbid intelligence in dementia. British Journal of Clinical Psychology. 1986; 25(Pt. 2):157–158. [PubMed: 3730655]

Patrick G, Gilbert J. Studies from the psychological laboratory of the University of Iowa: On the effects of loss of sleep. Psychological Review. 1896; 3:469–483.

Philibert I. Sleep loss and performance in residents and nonphysicians: A meta-analytic examination. Sleep. 2005; 28(11):1392–1402. [PubMed: 16335329]

Pilcher JJ, Band D, Odle-Dusseau HN, Muth ER. Human performance under sustained operations and acute sleep deprivation conditions: Toward a model of controlled attention. Aviation, Space, and Environmental Medicine. 2007; 78(Suppl 5):B15–B24.

Pilcher JJ, Huffcutt AI. Effects of sleep deprivation on performance: A meta-analysis. Sleep. 1996; 19(4):318–326. [PubMed: 8776790]

Price NJ, Maislin G, Powell JW, Ecker AJ, Szuba MP, Mallis MM, Dinges DF. Unobtrusive detection of drowsiness-induced PVT lapses using infrared retinal reflectance of slow eyelid closures. Sleep. 2003; 26(Abstract Suppl):A177.

Quigley N, Green JF, Morgan D, Idzikowski C, King DJ. The effect of sleep deprivation on memory and psychomotor function in healthy volunteers. Human Psychopharmacology. 2000; 15(3):171–177. [PubMed: 12404330]

Raven, J.; Raven, JC.; Court, JH. Manual for Raven's Progressive Matrices and Vocabulary Scales. Oxford, England: Oxford Psychologists Press; 1998.

Rosenthal R. Meta-analysis: A review. Psychosomatic Medicine. 1991; 53(3):247–271. [PubMed: 1882008]

Smith AP, Maben A. Effects of sleep deprivation, lunch, and personality on performance, mood, and cardiovascular function. Physiology & Behavior. 1993; 54(5):967–972. [PubMed: 8248391]

Smith ML, Glass GV. Meta-analysis of psychotherapy outcome studies. American Psychologist. 1977; 32(9):752–760. [PubMed: 921048]

Spreen, O.; Strauss, E. A compendium of neuropsychological tests: Administration, norms and commentary. New York, NY: Oxford University Press; 1991.

Stickgold R, Walker MP. Memory consolidation and reconsolidation: What is the role of sleep? Trends in Neurosciences. 2005; 28(8):408–415. [PubMed: 15979164]

Stroop JR. Studies of interference in serial verbal reactions. Journal of Experimental Psychology. 1935; 18(6):643–662.

Strube MJ, Hartmann DP. Meta-analysis: Techniques, applications, and functions. Journal of Consulting and Clinical Psychology. 1983; 51(1):14–27.

Sturm W, de Simone A, Krause BJ, Specht K, Hesselmann V, Radermacher I, Wilmes K. Functional anatomy of intrinsic alertness: Evidence for a fronto-parietal-thalamic-brainstem network in the right hemisphere. Neuropsychologia. 1999; 37(7):797–805. [PubMed: 10408647]

Sturm W, Willmes K. On the functional neuroanatomy of intrinsic and phasic alertness. NeuroImage. 2001; 14(1 Pt 2):S76–S84. [PubMed: 11373136]

Van Dongen HP, Baynard MD, Maislin G, Dinges DF. Systematic interindividual differences in neurobehavioral impairment from sleep loss: Evidence of trait-like differential vulnerability. Sleep. 2004; 27(3):423–433. [PubMed: 15164894]

Van Dongen, HP.; Dinges, DF. Circadian rhythm in sleepiness, alertness and performance. In: Kryger, MH.; Roth, T.; Dement, WC., editors. Principles and practice of sleep medicine. 4. Philadelphia, PA: Saunders; 2005. p. 435-443.

Watt KJ, O'Carroll RE. Evaluating methods for estimating premorbid intellectual ability in closed head injury. Journal of Neurology, Neurosurgery & Psychiatry. 1999; 66(4):474–479.

Wechsler, D. Wechsler Adult Intelligence Scale. 3. San Antonio, TX: Psychological Corporation; 1997a.

Wechsler, D. Wechsler Memory Scale. 3. San Antonio, TX: Psychological Corporation; 1997b.

Wilkinson RT. Interaction of lack of sleep with knowledge of results, repeated testing, and individual differences. Journal of Experimental Psychology. 1961; 62(3):263–271. [PubMed: 14007063]

Williams HL, Lubin A, Goodnow JJ. Impaired performance with acute sleep loss. Psychological Monographs. 1959; 73(14, Whole No. 484)
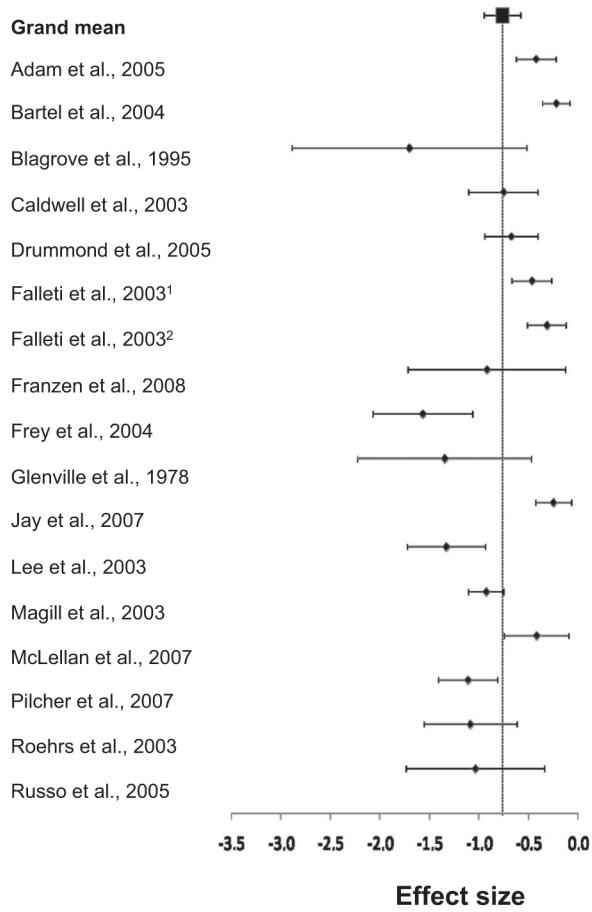
¹ Simple reaction time. ² Monitoring.

Figure 1.
Forest plots for a sample cognitive domain (lapses in simple attention tests). Effect sizes and 95% confidence intervals are plotted for the effect of short-term sleep deprivation on lapses in tests of simple attention. See the supplemental materials file for references to the studies cited here.
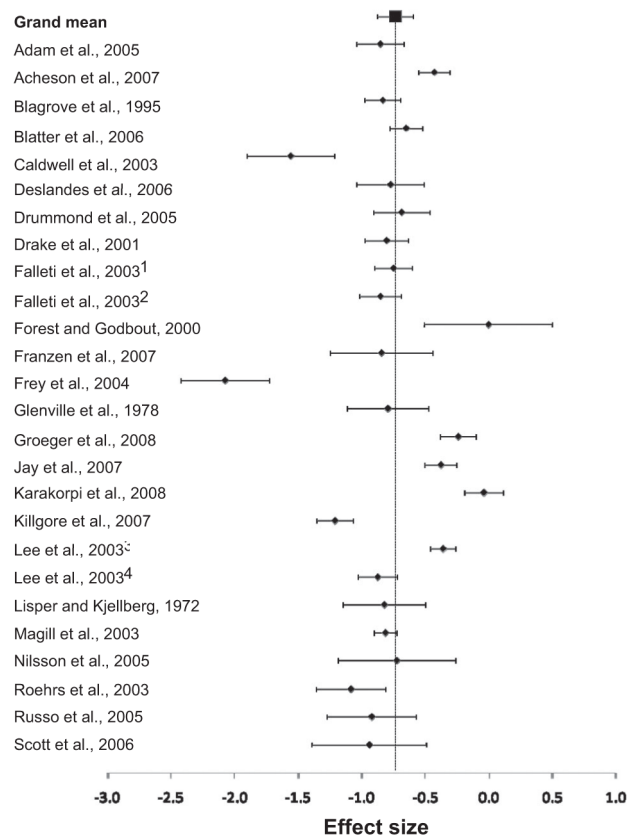
Figure 2.
Forest plots for a sample cognitive domain (reaction times in simple attention tests). Effect sizes and 95% confidence intervals are plotted for the effect of short-term sleep deprivation on reaction times in tests of simple attention. See the supplemental materials file for references to the studies cited here.
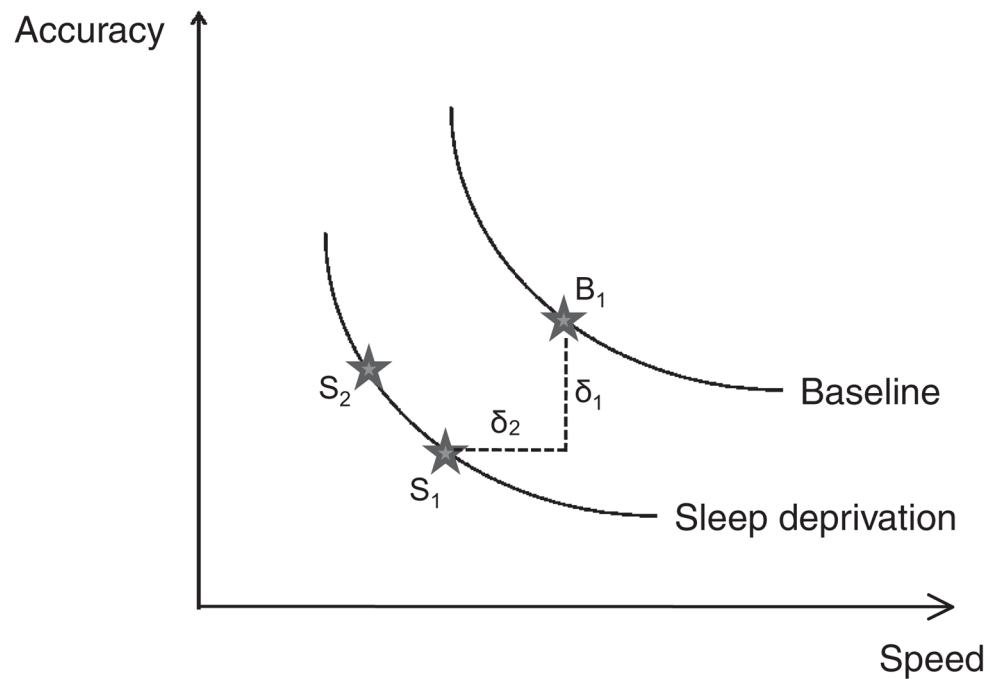
Figure 3.
Illustration of two possible ways in which sleep deprivation (SD) can affect speed and accuracy variables. Two sources of change may potentially occur following a period of total SD: a downward shift of the performance curve and a movement along the curve. In the case where $\delta_1 = \delta_2$ (i.e., there is a move to point S_1), no bias toward speed or accuracy occurs following SD. A movement along the curve (i.e., to S_2), however, would represent not just an overall decrement in performance but also a bias toward more accurate responding.

Table 1

List of Studies and Effect Sizes

Reference	Type	N	Time awake (hr)	Study quality	Cognitive test	Cognitive process	Effect size (Hedges's g)		
							Accuracy	Speed	Lapses
Acheson et al., 2007	WS	20	24	5.5	Stop task	CA	−0.931	0.086	
					SRT	SA		−0.850	—
					Two-choice reaction time test	CA	—	−0.926	
					Mathematical processing	PS	—	−0.777	
					Code substitution	PS	—	−0.693	
					Delayed matching to sample	WM	—	−0.623	
Adam et al., 2006[a]	WS	23	22	5.5	PVT	SA		−0.425	−0.424
Alhola et al., 2005	WS	10[b]	26	4	Visual episodic memory (short term)	STMc	0.093		
					DSST	PS	—	0.00	
					Cancellation	PS	—	0.244	
Bartel et al., 2004	WS	33	24–25[c]	3.5	SRT	SA	−0.829	−0.219	
					Complex reaction time test	CA	−0.435	−0.587	
					Sequential reaction time test (1)	CA	−0.590	−0.410	
					Sequential reaction time test (2)	CA	−0.917	−0.256	
Bell-McGinty et al., 2004[d]	WS	19	48	6	Nonverbal recognition task	WM	−1.60	0.372	
Binks et al., 1999	WS	61	34–36	6.5	Stroop test	CA	−0.511	−0.201	
					PASAT	PS	−0.211	—	
Blagrove and Akhurst, 2001	BS	61	29–35	5	Baddeley's Logical Reasoning	RE	−1.23		
Blagrove et al., 1995[e]	BS	14	26	4.5	Auditory vigilance test	SA		—	−1.70
		16			Finding embedded figures	PS	−0.996	—	
Blatter et al., 2006	WS	32	40	7	PVT	SA		−0.645	
Bray et al., 2004	WS	10[f]	24	5.5	Hopkins's Verbal Learning Test	STMc	−0.391		
					Digit span	WM	−0.079	—	
					Stroop test	CA	−0.183	—	
					Trail-making test	CA	—	−0.447	
Caldwell et al., 2003[g]	WS	16	28	5	PVT	SA		−1.555	−0.750

Reference	Type	N	Time awake (hr)	Study quality	Cognitive test	Cognitive process	Accuracy	Speed	Lapses
							Effect size (Hedges's g)		
Casagrande et al., 1997	WS	20	24	4	Letter cancellation task	PS	−0.510	−0.529	
Chee and Choo, 2004[d]	WS	14	24	6.5	SWM task	WM	−0.358	−0.629	
Chee et al., 2006[d]	WS	26	35	6.5	SWM task	WM	−1.16	−0.397	
Chee et al., 2008[d]	WS	24	24	6.5	Local/global task	CA	−1.04	−0.458	
Choo et al., 2005[d]	WS	14	24	5.5	n-back task	WM	−0.660	−0.573	
Chuah and Chee, 2008[d]	WS	28[f]	24	6	Visual short-term memory task	STMc	−1.26		
Chuah et al., 2006[d]	WS	27	24	6.5	Go/no-go task	CA	−0.337	—	
Corsi-Cabrera et al., 1996[h]	WS	9	40	3	Visual vigilance test	CA[i]	—	−2.241	
De Gennaro et al., 2001	WS	8	40	5	Letter cancellation	PS	0.00	−2.312	
Deslandes et al., 2006[h]	WS	10	24	4.5	SRT	SA	−0.770	−0.770	—
					Digit span	WM	−0.068	—	
					Stroop test	CA	0.098	0.060	
Drake et al., 2001	WS	10[f]	24	7	Paired-associates test	STMc	−0.699		
					PVT	SA		−0.682	—
Drummond et al., 1999[d]	WS	13	25	4	Serial subtraction task	PS	−0.457	—	
Drummond et al., 2000[d]	WS	13	35	6	Verbal memory (recall)	STMc	−0.582		
					Verbal memory (recognition)	STMg	−0.404		
Drummond et al., 2001[dj]	WS	13	35	6	Serial subtraction task	PS	−0.388	0.107	
					Verbal memory (recall)	STMc	−0.430		
					Verbal memory (recognition)	STMg	−0.619		
Drummond et al., 2005[d]	WS	20	36	6.5	PVT	SA		−0.800	−0.674
Drummond et al., 2006[d]	WS	38	48	5.5	Go/no-go task	CA	0.00	−0.717	
Falleti et al., 2003	WS	26	24	4	SRT	SA		−0.849	−0.466
					CRT	CA	−0.199	−1.02	
					Complex reaction time test	CA	−0.314	−0.529	
					Monitoring	SA		−0.748	−0.314
					One-back task	WM	−0.616	−0.405	
					Matching	CA	−0.851	−0.518	

Reference	Type	N	Time awake (hr)	Study quality	Cognitive test	Cognitive process	Effect size (Hedges's g)		
							Accuracy	Speed	Lapses
Fluck et al., 1998[k]	WS	6	24[c]	4.5	Associative learning	WM	−0.631	0.00	
					Logical memory (from WMS)	STMc	−0.119		
					Digit cancellation	PS	—	0.00	
					Digit span	WM	—	0.00	
					DSST	PS	−0.337	—	
					PASAT	PS	—	−0.238	
					Trail-making test	CA	—	−0.735	
Forest and Godbout, 2000	BS	18	25.5	4.5	Continuous performance test	CA	—	0.00	
					Target detection	CA	—	0.00	
					SRT	SA		0.00	
Franzen et al., 2008	BS	29	24	7	PVT	SA	−0.840	−0.920	
Frey et al., 2004	WS	25	43	5	PVT	SA		−2.07	−1.569
					Wilkinson Four-Choice Reaction Time Test	CA	−1.10	−1.23	
					Two-column addition test	PS		−0.225	
					Digit recall	WM	−0.717	0.00	
					Reaction time test	WM	−0.681	−1.418	
Glenville et al., 1978[l]	WS	8	30	4.5	Vigilance test	SA	—		−1.35
					SRT	SA		−0.79	
					CRT	CA	0.00	−0.79	
					Digit span	WM	0.00	—	
Gosselin et al., 2005[h]	BS	24	36	6	Auditory oddball task	CA	−0.919	0.017	
Groeger et al., 2008[m]	WS	24	40	6	Spatial–verbal n-back	CA	−0.737[n]	—	
					Sustained attention to response task	SA	—		−0.632
					Serial addition	PS	—	−0.530	
					DSST	PS	—	0.00	
					SRT	SA	—	−0.236	
					Serial reaction time test	CA	—	−0.272	
Habeck et al., 2004[d]	WS	18	48	6	SWM task	WM	−0.939	−1.50	

Reference	Type	N	Time awake (hr)	Study quality	Cognitive test	Cognitive process	Effect size (Hedges's g)		
							Accuracy	Speed	Lapses
Harrison and Horne, 1998	BS	20	34	6	Hayling test	CA	−0.594	−0.837	
Harrison and Horne, 2000	BS	20	35	6	Temporal memory for faces	STMc	−0.008	—	
					Self-ordered pointing	PS	0.00	—	
Heuer et al., 2005	WS	17	24	3	Simon test	CA	—	0.00	
					Stroop test	CA	—	−0.524	
Horne, 1988	BS	24	32	4	Nonverbal planning	CA	—	−0.726	
Hsieh et al., 2007[h]	WS	16	24	7	Flanker task	CA	−0.262	−0.085	
Jay et al., 2007[o]	WS	20	40	7	PVT	SA		−0.343	−0.248
Karakorpi et al., 2006	WS	21[f]	40	4	SRT	SA		−0.036	—
					Two-choice reaction time	CA	—	0.652	
Killgore et al., 2007	WS	54	23	5	PVT	SA		−1.209	—
Kim et al., 2001	WS	18	24	5.5	Digit span	WM	−0.213	—	
Lee et al., 2003[h,j]	WS	20	28	4.5	Vigilance task	SA		−0.871	−1.33
					SRT	SA		−0.355	
					Cognitrone (symbol search)	PS	0.467	0.030	
Lim et al., 2007[dq]	WS	19	24	6	SWM task	WM	−1.52	−0.548	
Lim et al., in preparation[d]	WS	24	24	7	Selective attention test	CA	−1.15	−0.271	
Linde and Bergstrom, 1992	BS	16	24	4.5	Raven's Progressive Matrices	RE	−0.921		
					Baddeley's logical reasoning	RE	−1.21		
					Digit span	WM	0.00	—	
Linde et al., 1999	BS	24	33	3	Coding	PS	−0.691	−1.37	
					Raven's Progressive Matrices	RE	0.201		
Lisper and Kjellberg, 1972[l]	WS	8	24	4	SRT	SA		−0.817	
Luber et al., 2008[r]	WS	8	48	4.5	SWM task	WM	—	−1.43	
Magill et al., 2003	WS	76[f]	30	6	Visual scanning	PS	—	0.00	
					One-back task	WM	0.00	−0.387	
					Baddeley's logical reasoning	RE	0.00	0.00	
					Mathematical processing	PS	0.00	−0.812	
					Stroop test	CA	0.00	0.00	

Reference	Type	N	Time awake (hr)	Study quality	Cognitive test	Cognitive process	Effect size (Hedges's *g*)		
							Accuracy	Speed	Lapses
					Four-choice reaction time test	CA	—	0.00	
					Visual vigilance test	SA		−0.809	−0.928
					Trails (B) task	CA	—	0.00	
Mander et al., 2008[ds]	WS	7	34–36	4.5	Posner cueing paradigm	CA	−2.31	−0.225	
McCarthy and Waters, 1997[h]	WS	35	36	4.5	Stroop test[t]	CA	−0.223	0.117	
					Memory search	STMg	−0.342	0.00	
					Baddeley's Logical Reasoning	RE	−0.236	0.00	
McLellan et al., 2007[u]	WS	10[f]	21	3	Vigilance test	SA		—	−0.419
McMorris et al., 2007	WS	10[f]	36	3.5	Digit span	WM	0.224	—	
					CRT	CA	—	0.370	
Mu et al., 2005[d]	WS	33	30	5.5	SWM task	WM	−1.01	−0.312	
Murphy et al., 2006[h]	WS	17	19	5.5	Flanker test	CA	0.122	—	
Nilsson et al., 2005	BS	22	32	7	Verbal learning	STMc	0.886		
					Visuospatial working memory	WM	−0.150	—	
					SRT	SA		−1.21	
O'Hanlon and Vermeeren, 1988	WS	8[f]	26	4.5	Vigilance test	CA[i]	−1.00	—	—
Pilcher et al., 2007	WS	38	28	6.5	PVT	SA		—	−1.11
					Code substitution	PS	—	0.00	
					SWM task	WM	—	−1.36	
					Continuous performance test	CA	—	−0.306	
Quigley et al., 2000	BS	26	24	6	Digit span	WM	0.00	—	
					Word recall	STMc	0.00		
					Word recognition	STMg	0.00		
					DSST	PS	0.00	—	
					Prose recall	STMc	0.00		
Roehrs et al., 2003	WS	16	24	5	Paired-associates test	STMc	−0.589		
					PVT	SA		−1.082	−1.089
Russo et al., 2005	WS	8	26.5	3.5	PVT	SA		−0.918	−1.04

Reference	Type	N	Time awake (hr)	Study quality	Cognitive test	Cognitive process	Effect size (Hedges's g)		
							Accuracy	Speed	Lapses
Scott et al., 2006[v]	WS	6	30	4	SRT	SA	—	-0.936	—
					Two-choice reaction time test	CA	—	-0.052	
					Number cancellation task	PS	—	-0.412	
Smith and Maben, 1993	BS	21	27–32	5	Baddeley's Logical Reasoning	RE	0.00		
					One-back task	WM	-1.745	—	
Thomas et al., 2000[w]	WS	17	24	6	Serial addition/subtraction task	WM	-0.536	-0.361	
Tsai et al., 2005[h]	WS	16	26	6.5	Flanker test	CA	-0.759	-0.689	
Turner et al., 2007	WS	40	42	6	Continuous paired associates (variant)	WM	-1.034	—	

Note. For some imaging (i.e., positron emission tomography, functional magnetic resonance imaging, and electroencephalograph) studies, reaction times are aggregated and reported with only correct responses; this may have led in some cases to an underestimate of the true effect size. Dashes indicate data not reported. WS = within subjects; CA = complex attention; SA = simple attention; PS = processing speed; WM = working memory; PVT = Psychomotor Vigilance Test; STMc = short-term memory recall; DSST = Digit Symbol Substitution Test; SRT = simple reaction time; PASAT = Paced Auditory Serial Addition Test; BS = between subjects; RE = reasoning; SWM = Sternberg working memory; STMg = short-term memory recognition; CRT = choice reaction time; WMS = Wechsler Memory Scale.

[a]Subjects included 11 older (61–70 years) males.

[b]Only data from nonhormone therapy users were entered into the analysis.

[c]Subjects were junior doctors on call and received a small amount of sleep during the night.

[d]Tasks were performed while subjects underwent functional magnetic resonance imaging scanning.

[e]Data from the "second sleep deprivation study" were used for effect size calculation.

[f]Only data from the placebo group or session were used.

[g]Data were collected from subjects in two postures (standing and sitting); these were combined in this analysis.

[h]Tasks were performed while subjects underwent electroencephalograph monitoring.

[i]Although labeled as vigilance tests, these were described as containing an element of selectivity and thus categorized under "complex attention."

[j]Although the paradigm in this study was a divided attention task, data from each task were reported independently in both states, thus allowing for a comparison of each cognitive process.

[k]Results reported here were calculated from Experiment 2 ("Six junior doctors tested twice …").

[l]The *F* value for the 50th percentile (i.e., median reaction time) change after sleep deprivation was used in calculating the effect size.

[m] Data for this study were reported relative to the melatonin midpoint for each subject. Effect sizes were calculated on the basis of change from baseline performance to performance at the melatonin midpoint, which was around 0400 hr for each subject.

[n] Results were pooled across all six conditions of the spatial and verbal *n*-back tasks.

[o] Subjects in this experiment were randomized into 6-hr and 9-hr recovery sleep conditions; as we were not concerned with these data, outcomes across both groups were combined for the baseline and sleep deprivation nights.

[p] Data from "pre-D1" and "post-D1" were used.

[q] Subjects underwent two sessions of 24-hr sleep deprivation; data across these sessions were aggregated.

[r] Data from the control experiment were used.

[s] Results for valid, neutral, and invalid cues were pooled.

[t] Outcomes for the Stroop test were aggregated over three forms of the test.

[u] Data compared were from Day 1 (control) and Day 3, session 1.

[v] Only data from the baseline and not the exercising condition were used.

[w] Although subjects were awake for 85 hr during this protocol, only data from the test at 24-hr sleep deprivation were reported here. Tasks were performed while subjects underwent positron emission tomography scanning.

Table 2

Combined Effect Sizes by Domain and Outcome Variable

Outcome variable	Combined effect size	Variance	SE	95% CI		Q	df	I²
				LL	UL			
Simple attention								
Lapses	−0.762**	0.009	0.095	−0.948	−0.576	112.18	16	61.6
Reaction time	−0.732**	0.005	0.072	−0.874	−0.590	97.04	25	54.1
Complex attention								
Accuracy	−0.479**	0.007	0.082	−0.640	−0.318	56.79	24	31.7
Reaction time	−0.312**	0.003	0.059	−0.429	−0.197	192.57	36	53.5
Processing speed								
Accuracy	−0.245	0.017	0.130	−0.500	0.010	72.99	11	52.1
Reaction time	−0.302**	0.007	0.083	−0.464	−0.140	194.77	19	62.4
Working memory								
Accuracy	−0.555**	0.009	0.095	−0.741	−0.368	113.79	25	55.4
Reaction time	−0.515**	0.009	0.097	−0.704	−0.326	92.95	16	66.3
Short-term memory								
Recall	−0.383*	0.018	0.135	−0.647	−0.118	37.85	11	55.1
Recognition	−0.378*	0.016	0.125	−0.624	−0.132	13.91	4	13.9
Reasoning								
Accuracy	−0.125	0.005	0.072	−0.268	0.016	14.59	11	5.4

Note. CI = confidence interval; LL = lower limit; UL = upper limit.

*
$p < .01$.

**
$p < .001$.

Table 3

Analysis of Variance Comparing Average Effect Sizes Within Outcome Variable Type and Across Cognitive Domains

Outcome variable	Combined effect size	Variance	Q	df
Accuracy and lapses	-0.407 [**]	0.001	33.94	6
Reaction time	-0.450 [**]	0.001	25.63	3

[**] $p < .001$.

Table 4

Analysis of Variance Comparing Average Effect Sizes Within Cognitive Domains and Across Outcome Variable Type

Cognitive domain	z
Simple attention	0.256
Complex attention	1.645
Working memory	0.555
Processing speed	0.292

Note. None of these differences were significant at the $p < .05$ level.