A new metric for quantifying performance impairment on the psychomotor vigilance test

SRINIVASAN RAJARAMAN[1], SRIDHAR RAMAKRISHNAN[1], DAVID THORSLEY[1], NANCY J. WESENSTEN[2], THOMAS J. BALKIN[2] and JAQUES REIFMAN[1]

[1]DoD Biotechnology High-Performance Computing Software Applications Institute, Telemedicine and Advanced Technology Research Center, US Army Medical Research and Materiel Command, Fort Detrick, MD, USA and [2]Behavioral Biology Branch, Center for Military Psychiatry and Neuroscience, Walter Reed Army Institute of Research, Silver Spring, MD, USA

Keywords
cognitive performance, effect size, Jensen–Shannon divergence, psychomotor vigilance test, sleep deprivation, two-process model

Correspondence
Jaques Reifman, PhD, Senior Research Scientist, DoD Biotechnology HPC Software Applications Institute, Telemedicine and Advanced Technology Research Center, US Army Medical Research and Materiel Command, ATTN: MCMR-TT, 504 Scott Street, Fort Detrick, MD 21702, USA.
Tel.: 301-619-7915;
fax: 301-619-1983;
e-mail: jaques.reifman@us.army.mil

Accepted in revised form 8 February 2012; received 27 September 2011

DOI: 10.1111/j.1365-2869.2012.01008.x

SUMMARY

We have developed a new psychomotor vigilance test (PVT) metric for quantifying the effects of sleep loss on performance impairment. The new metric quantifies performance impairment by estimating the probability density of response times (RTs) in a PVT session, and then considering deviations of the density relative to that of a baseline-session density. Results from a controlled laboratory study involving 12 healthy adults subjected to 85 h of extended wakefulness, followed by 12 h of recovery sleep, revealed that the group performance variability based on the new metric remained relatively uniform throughout wakefulness. In contrast, the variability of PVT lapses, mean RT, median RT and (to a lesser extent) mean speed showed strong time-of-day effects, with the PVT lapse variability changing with time of day depending on the selected threshold. Our analysis suggests that the new metric captures more effectively the homeostatic and circadian process underlying sleep regulation than the other metrics, both directly in terms of larger effect sizes (4–61% larger) and indirectly through improved fits to the two-process model (9–67% larger coefficient of determination). Although the trend of the mean speed results followed those of the new metric, we found that mean speed yields significantly smaller (~50%) intersubject performance variance than the other metrics. Based on these findings, and that the new metric considers performance changes based on the entire set of responses relative to a baseline, we conclude that it provides a number of potential advantages over the traditional PVT metrics.

INTRODUCTION

The psychomotor vigilance test (PVT) (Dinges and Powell, 1985) is a well-validated, widely used tool for assessing neurobehavioral impairment due to both total sleep deprivation and chronic sleep restriction (Dorrian *et al.*, 2005). Moreover, the PVT is not influenced by an individual's aptitude; its results are immune to practice effects, and track accurately the interaction between the homeostatic drive for sleep and the circadian rhythm of alertness. Therefore, several metrics, including mean and median response times (RTs), mean and median speeds (i.e. the reciprocal of RT) and threshold-based lapses (e.g. number of RTs >500 ms), have been proposed and used to quantify PVT performance.

However, despite nearly three decades of sleep loss research using the PVT, there has been little attempt to understand the merits and demerits of these PVT-derived metrics. In particular, it has not been considered that the existing metrics may capture the information generated in a PVT session incompletely and, accordingly, only partially reflect the neurobehavioral state of the individual being tested. For example, the number of PVT lapses above a 500-ms threshold, which is considered to be a well-validated PVT performance metric, reflects only the proportion of the total responses that are >500 ms without providing any information about their relative frequency. Another related problem is that PVT lapses map a continuous variable (i.e. RT) into a discrete, binary variable representation (i.e. lapse/no lapse),

potentially losing useful information in this process. In contrast, while RT summary statistics, including mean and median RTs and mean and median speeds, are inherently more comprehensive at representing the information contained in RTs than PVT lapses, they fail to capture the information content of the shape of the RT density. Another limitation of these PVT metrics is their inherent insensitivity to baseline performance levels. That is, these metrics do not directly assess an individual's performance relative to the individual's well-rested performance, necessarily assuming that baseline performance is invariant across individuals.

The objective of this paper is twofold: (1) to identify and characterize ambiguities and gaps in inferences generated by the existing PVT metrics and (2) to propose a new PVT performance metric that addresses some of the aforementioned limitations in capturing performance impairment due to total sleep loss.

METHODS

The proposed metric attempts to quantify PVT performance more comprehensively than the existing metrics in two important ways: (1) by accounting for the density of the measured RTs in each PVT session and (2) by inherently considering deviations relative to baseline performance levels. This was achieved by estimating the probability density function (PDF) of RTs for each PVT session, establishing a baseline PVT session, and quantifying the performance on a PVT session as the dissimilarity between its RT PDF and that of the baseline session.

Empirical estimation of RT PDF

First-principles techniques have been proposed to describe the RT PDFs for simple reaction tasks (Van Zandt, 2000). However, these techniques require the censoring of responses that are slower and faster than preset thresholds before the density can be estimated with the remaining RT data. Therefore, we adopted an empirical approach to estimate the density of RTs measured in a PVT session without discarding either slow or fast responses. A detailed description of the empirical density estimation approach is provided in the Appendix, Section I.

Identification of the baseline PVT session

Currently, there is no consensus in the definition of a baseline PVT session. Because sleep loss increases both fast and slow responses (Dorrian *et al.*, 2005) which, in turn, increases the skewness in RT densities, we defined the baseline PVT session as the test session in which both tails of the RT density were balanced or, equivalently, the session in which the RT density was as close as possible to a normal density. This hypothesis is supported by recent work by Holden *et al.* (2009), who showed that the overall response latency is a result of interaction between subsidiary components of human cognition. During baseline conditions, the level of interaction is the least, i.e. the subsidiary components vary as independent, random variables, yielding response latency that follows a normal density.

The new PVT metric

We propose a new PVT metric, termed the response time divergence (RTD), which defines an individual's performance on a PVT session T as the dissimilarity between an individual's entire RT density at that session, $p_T(t)$, and the RT density at baseline, $p_{BL}(t)$. Mathematically, the proposed metric is defined as follows:

$$\mathrm{RTD}(T) = \mathrm{sgn}[p_T(t), p_{BL}(t)] \frac{N_T}{\sqrt{\ln(2)}} \sqrt{\mathrm{JSD}[p_T(t), p_{BL}(t)]}, \quad (1)$$

where N_T represents the total number of responses observed in session T, $\mathrm{sgn}[p_T(t), p_{BL}(t)]$ defines the direction of change in performance from baseline to session T and $\mathrm{JSD}[p_T(t), p_{BL}(t)]$ represents the Jensen–Shannon divergence (Lin, 1991) between $p_T(t)$ and $p_{BL}(t)$. In eqn (1), as $p_T(t)$ diverges from $p_{BL}(t)$, the metric increases, attaining a maximum value N_T. Conversely, as the two densities converge, RTD(T) approaches zero (Appendix, Section II).

Measures for comparing and contrasting PVT performance metrics

We compared and contrasted the RTD metric against existing PVT metrics in their ability to capture the homeostatic and circadian processes underlying sleep regulation by: (1) performing effect size analysis and (2) assessing the goodness-of-fit of each PVT metric to Borbély's two-process model output. For the latter, we compared the coefficient of determination (R^2; Zar, 1999) and the degree of whiteness (randomness) of the residual error, i.e. the difference between the individualized two-process model fit and the PVT metric (Appendix, Section III).

Laboratory study data

We used PVT data obtained from a laboratory study involving 12 subjects (mean age = 24.9 years, range = 19–39 years) who were kept awake continuously for 85 h (Wesensten *et al.*, 2005). These subjects were administered PVTs once every 2 h starting at 08:00 h on day 1 (00:00–24:00 h) and extending through 18:00 h on day 4, for a total of 42 PVT sessions. At 20:00 h on day 4, subjects initiated recovery sleep for 12 h. Bi-hourly PVTs were resumed at 10:00 h on day 5 and ended at 16:00 h. The study was approved by the Walter Reed Army Institute of Research Human Use Committee (Silver Spring, MD, USA) and the US Army Medical Research and Materiel Command Human Subjects Review

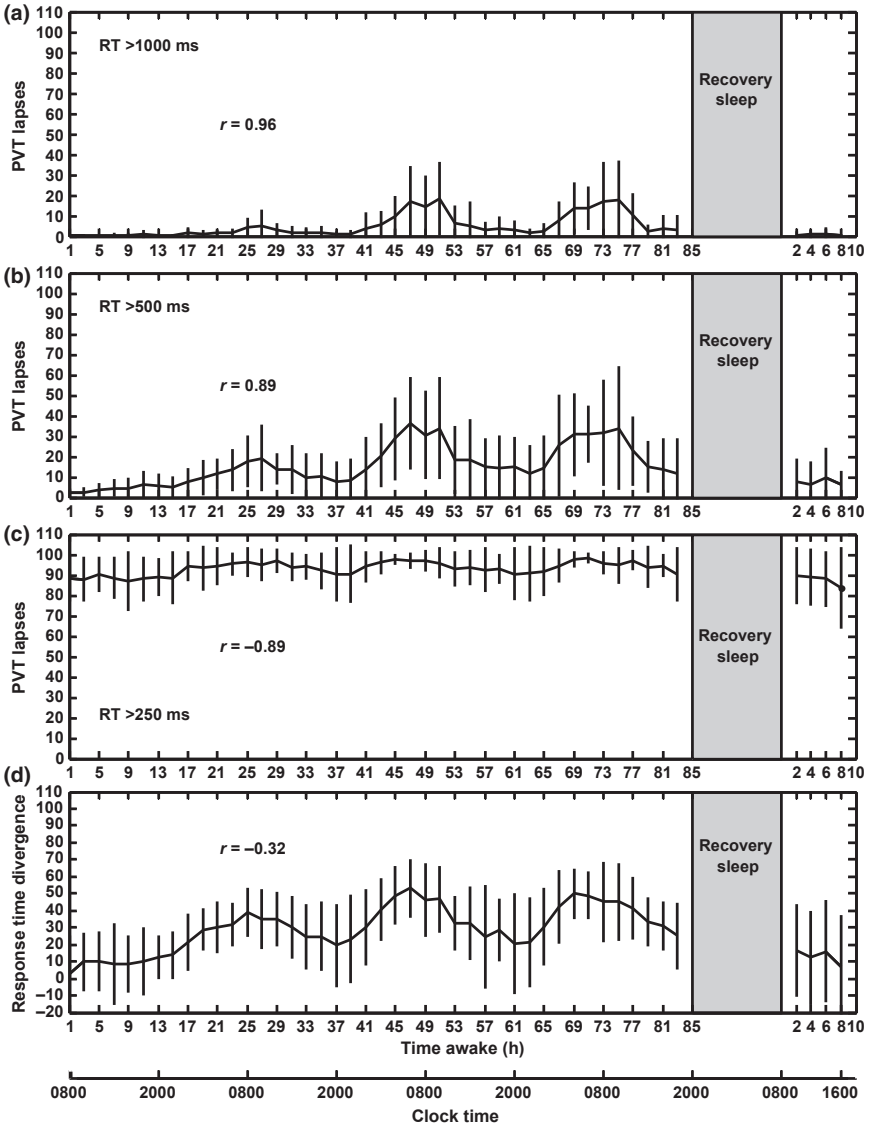

Figure 1. Mean and standard deviation (SD) of psychomotor vigilance test (PVT) lapse measurements (n = 12) for three different thresholds: (a) 1000 ms, (b) 500 ms and (c) 250 ms, and (d) for the response time divergence (RTD) metric. We used Pearson's correlation coefficient (r) to compute the correlation between the temporal profiles of the mean and SD as a function of wakefulness. The shaded region in each plot represents the 12-h recovery sleep administered to the subjects immediately after 85 h of continuous sleep deprivation.

Board (Fort Detrick, MD, USA). Written informed consent was obtained from all subjects prior to their participation.

RESULTS

We first analyzed the temporal profiles of the mean and standard deviation (SD) across the 12 subjects for each metric. Fig. 1a shows that the mean and SD of PVT lapses for RT >1000 ms increased during the early morning and decreased between late afternoon and early evening on all days of sleep deprivation. The temporal profiles of these statistics over the 12 subjects was highly correlated (Pearson's correlation coefficient r = 0.96), exhibiting an almost perfect linear relationship. We found a similar correlation for RT >500 ms (Fig. 1b), but to a lesser extent (r = 0.89) than that observed for RT >1000 ms. Surprisingly, for RT >250 ms, the temporal profiles of the mean and SD were correlated negatively (r = −0.89; Fig. 1c), i.e. performance variability was lower during the early morning hours than during the early evening hours, contrary to the time-of-day effects shown in Fig. 1a,b. We also found the correlations of these statistics for the mean RT, median RT and mean speed to fall on both sides of the spectrum. While for mean and median RT the group mean and SD were correlated highly, with r = 0.97 and 0.92, respectively, for mean speed they were correlated modestly negatively (r = −0.53).

This suggests that the existing metrics, and PVT lapses in particular, while directly not accounting for, are influenced by changes in the RT densities. Indeed, we found that the RT density may change significantly across wakefulness and

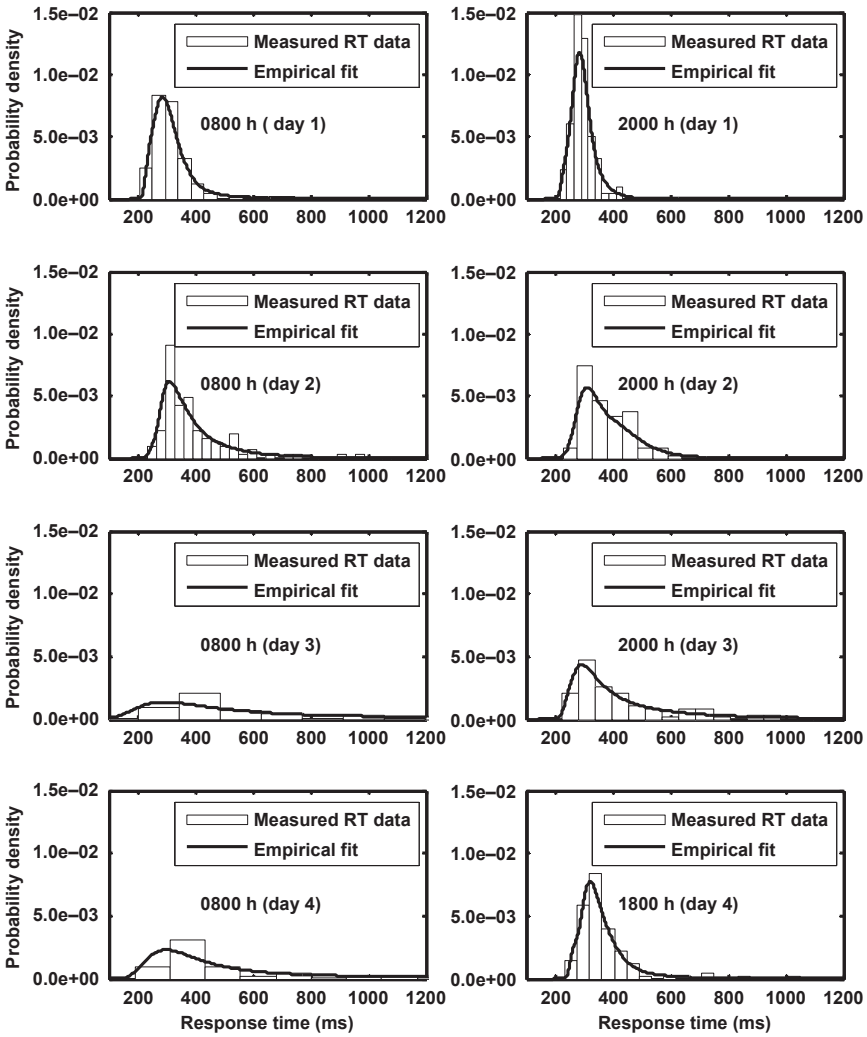

Figure 2. Histograms of measured response time (RT) data (stacked bars) observed every 12 h, starting at 08:00 h on day 1 and extending to 18:00 h on day 4 of the total sleep deprivation routine, and their corresponding empirical probability density estimates (solid lines) for a subject with an average sleep-loss phenotype (subject 12).

time of day. For example, Fig. 2 shows histograms of the measured RT data and the corresponding density estimates computed empirically at eight distinct time-points across wakefulness for a subject with an 'average' sleep-loss phenotype (subject 12; Table 2). Qualitatively, moving from days 1–4 (Fig. 2, left), the empirical density fits at 08:00 h reflected an increase in, or rather an elongation of, the right tail of the density, reflecting a systematic increase in the relative frequency of lapses with sleep deprivation. A similar pattern was observed with the empirical density fits at ~20:00 h on days 1–4 (Fig. 2, right), but to a lesser extent than those observed at 08:00 h on the corresponding days, reflecting the influence of circadian effects on performance. In fact, we also found that even the baseline RT densities of the 12 subjects were significantly different in 82% (54 of 66) of the possible distinct pairwise tests (two-sample Kolmogorov–Smirnov test for $P < 0.05$).

Fig. 1d shows the mean and SD of the RTD metric for the 12 subjects. Unlike the PVT lapses shown in Fig. 1a–c, the RTD metric did not express PVT performance variability unequally at different times across wakefulness, indicated by a relative lack of a correlation between mean and SD for this metric ($r = -0.32$). Our analysis suggested that by estimating performance relative to an individual's baseline, the RTD attenuates the large swings in intersubject variability across different phases of the circadian rhythm observed in PVT lapses (Appendix, Section IV).

We also computed the RTD for specific portions of the RT density (i.e. for RT >1000, 500 and 250 ms) and compared them with PVT lapses (Fig. 1a–c). We found that regardless of the chosen threshold, and unlike PVT lapses, for any specific portion of the RT density the RTD metric yielded approximately uniform intersubject variability, which was independent of time of day (Appendix, Section V).

To compare and contrast the PVT metrics in terms of their ability to capture the homeostatic and circadian processes underlying sleep regulation, we first performed group (within-subject) effect size analysis. We assessed the group effect

Table 1 Effect size scores between group performance at baseline (08:00–18:00 h, day 1) and those at 08:00–18:00 h on days 2–4, for the response time divergence (RTD), psychomotor vigilance test (PVT) lapses (500 ms), mean response time (RT), median RT and mean speed metrics. The entries inside the parentheses denote the non-parametric 95% bootstrap confidence intervals (CIs) of the effect size scores, where the 95% CIs were based on 1000 bootstrap samples

	Effect size (95% CI)				
Time of day (h)	*RTD*	*PVT lapses (500 ms)*	*Mean RT (ms)*	*Median RT (ms)*	*Mean speed (s^{-1})*
Day 2					
08:00	1.25 (0.89, 1.54)	1.20 (0.94, 1.37)	1.22 (1.03, 1.35)	1.34 (1.09, 1.51)	1.32 (1.14, 1.46)
10:00	1.32 (0.78, 1.71)	1.13 (0.91, 1.25)	0.88 (0.69, 1.03)	1.14 (0.79, 1.38)	1.30 (1.10, 1.47)
12:00	1.60 (0.91, 1.92)	1.47 (0.96, 1.76)	1.50 (1.05, 1.75)	1.74 (1.23, 2.01)	1.69 (1.32, 1.94)
14:00	1.12 (0.83, 1.52)	0.90 (0.63, 1.10)	1.15 (0.81, 1.38)	1.37 (0.99, 1.65)	1.70 (1.12, 2.05)
16:00	0.88 (0.58, 1.23)	0.59 (0.38, 0.75)	0.83 (0.61, 1.95)	0.96 (0.58, 1.21)	0.85 (0.64, 1.04)
18:00	0.99 (0.39, 1.16)	0.72 (0.35, 0.96)	1.06 (0.40, 1.36)	0.86 (0.52, 1.12)	1.06 (0.68, 1.33)
Day 3					
08:00	1.51 (1.04, 1.80)	1.33 (1.16, 1.46)	0.85 (0.72, 0.96)	0.93 (0.62, 1.15)	1.58 (1.45, 1.77)
10:00	1.56 (1.01, 2.04)	1.31 (1.17, 1.44)	0.75 (0.63, 0.86)	0.67 (0.47, 0.89)	1.46 (1.31, 1.61)
12:00	1.25 (0.39, 2.64)	0.93 (0.75, 1.07)	0.72 (0.61, 0.81)	1.02 (0.72, 1.21)	1.17 (0.94, 1.35)
14:00	0.82 (0.46, 1.39)	0.68 (0.55, 0.80)	0.46 (0.38, 0.56)	0.68 (0.50, 0.87)	0.97 (0.79, 1.11)
16:00	0.63 (0.38, 1.21)	0.76 (0.58, 0.89)	0.77 (0.61, 0.88)	0.66 (0.53, 0.80)	0.68 (0.56, 0.77)
18:00	0.94 (0.32, 0.83)	0.58 (0.40, 0.73)	0.58 (0.35, 0.77)	0.52 (0.41, 0.67)	0.76 (0.58, 0.92)
Day 4					
08:00	1.36 (0.99, 1.82)	1.15 (1.00, 1.21)	0.65 (0.56, 0.72)	0.70 (0.56, 0.86)	1.33 (1.20, 1.45)
10:00	1.34 (0.82, 1.38)	1.09 (0.98, 1.17)	0.77 (0.68, 0.84)	0.82 (0.63, 0.93)	1.20 (1.11, 1.32)
12:00	1.59 (1.09, 2.00)	1.15 (0.95, 1.31)	0.90 (0.74, 1.03)	1.24 (1.02, 1.44)	1.59 (1.31, 1.80)
14:00	0.86 (0.49, 0.95)	0.94 (0.67, 1.14)	1.03 (0.75, 1.18)	0.80 (0.66, 0.94)	0.85 (0.68, 0.99)
16:00	1.16 (0.70, 1.58)	0.79 (0.60, 0.93)	0.73 (0.58, 0.84)	1.31 (0.86, 1.48)	1.15 (0.89, 1.28)
18:00	0.80 (0.25, 0.81)	0.47 (0.30, 0.61)	0.45 (0.29, 0.61)	0.80 (0.54, 0.99)	0.64 (0.48, 0.86)
Overall					
08:00	1.37	1.15	0.79	0.91	1.39
10:00	1.37	1.13	0.78	0.82	1.29
12:00	1.48	1.06	0.81	1.23	1.41
14:00	0.89	0.77	0.59	0.82	0.95
16:00	0.81	0.72	0.77	0.79	0.79
18:00	0.89	0.54	0.53	0.63	0.75

size as a function of time of day (i.e. the circadian phase) by comparing performance on day 1 (i.e. the baseline day) with performance on days 2, 3 and 4 for each session between 08:00 and 18:00 h using the effect size d_{bp} in eqn (A5) in the Appendix. The results suggest that across all sessions between 08:00 and 18:00 h on days 2–4, the RTD metric scored a larger effect size than those of PVT lapses, mean RT, median RT and mean speed on 89, 78, 61 and 55% of the cases, respectively (Table 1). The table also shows that, overall, for each of the six sessions averaged across days 2–4 using the effect size $\bar{d}$ in eqn (A6) in the Appendix, the RTD metric scored a larger effect size than those of PVT lapses, mean RT, median RT and mean speed on 100, 100, 100 and 67% of the cases, respectively. When we combined the effect size scores for each metric across the six sessions, we found that the RTD metric scored the largest average effect size (1.14), ranging from 4 to 61% larger than those of the other four metrics, followed by mean speed, PVT lapses, median RT and mean RT.

In addition, we used the goodness-of-fit between each metric and the corresponding individualized two-process model output (Rajaraman *et al.*, 2008, 2009) to assess indirectly the ability of the metrics to capture the homeostatic and circadian processes underlying sleep loss. Figs 3, 4 and 5 show the PVT lapses for RT >500 ms and the RTD metric for three subjects with resilient (subject 9), vulnerable (subject 1) and average (subject 12) sleep-loss phenotypes, respectively, and their corresponding individualized two-process model fits (dashed line, both panels). The results for the resilient sleep-loss phenotype (Fig. 3) suggest that, in terms of R^2, the variance explained by the two-process model for the RTD metric was 8% larger than that for the PVT lapses, and in terms of *P*-value for Bartels's test of randomness of the residual error in the model fit, the results suggest that the RTD metric ($P = 0.12$) provided a superior goodness-of-fit to the two-process model of sleep regulation than the PVT-lapse metric ($P = 0.01$), which could not be characterized adequately by the two-process model. When we fitted the RTD and PVT-lapse data for the subjects with vulnerable (Fig. 4) and average (Fig. 5) sleep-loss phenotypes, the RTD also explained the variance (R^2) (66 and 17%, respectively) more clearly. For the *P*-value analysis, the results for the vulnerable subject (Fig. 4) suggest that the RTD metric provided a significantly better goodness-of-fit to the two-process model than lapses ($P = 0.52$ versus $P = 0.06$) and the results for the average subject (Fig. 5) suggest that both

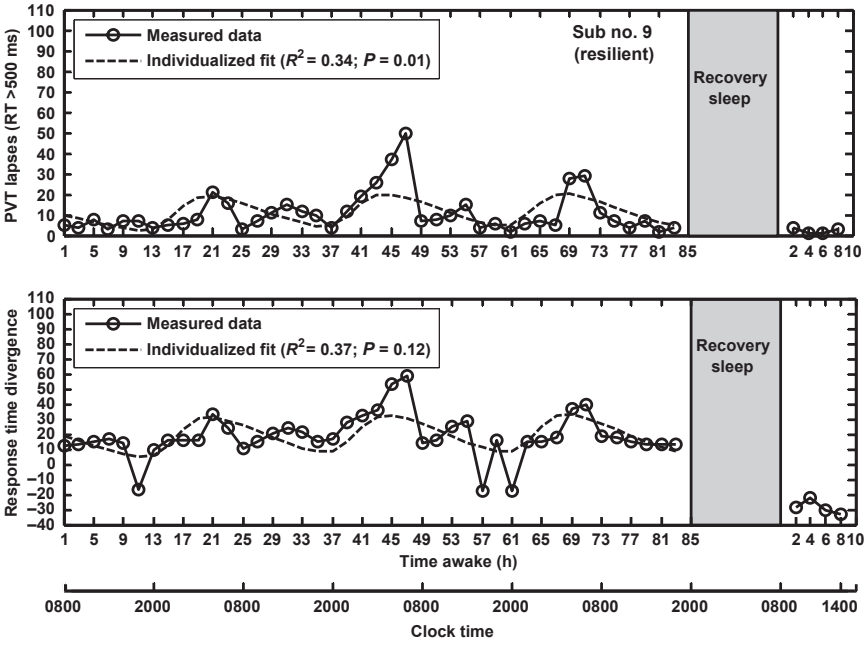

Figure 3. Psychomotor vigilance test (PVT) performance of a resilient individual (subject 9) over 82 h of continuous wakefulness measured once every 2 h in terms of PVT lapses (open circles, top) and the response time divergence (RTD) metric (open circles, bottom). The dashed line in each panel represents the individualized two-process model fit to the corresponding performance metric. We used the coefficient of determination (R^2) and the *P*-values based on Bartels's test of randomness of the residual error to quantify the goodness-of-fit of the two-process model to PVT lapses and the RTD metric. The shaded region in each plot represents the 12-h recovery sleep administered to the subjects immediately after 85 h of total sleep deprivation.

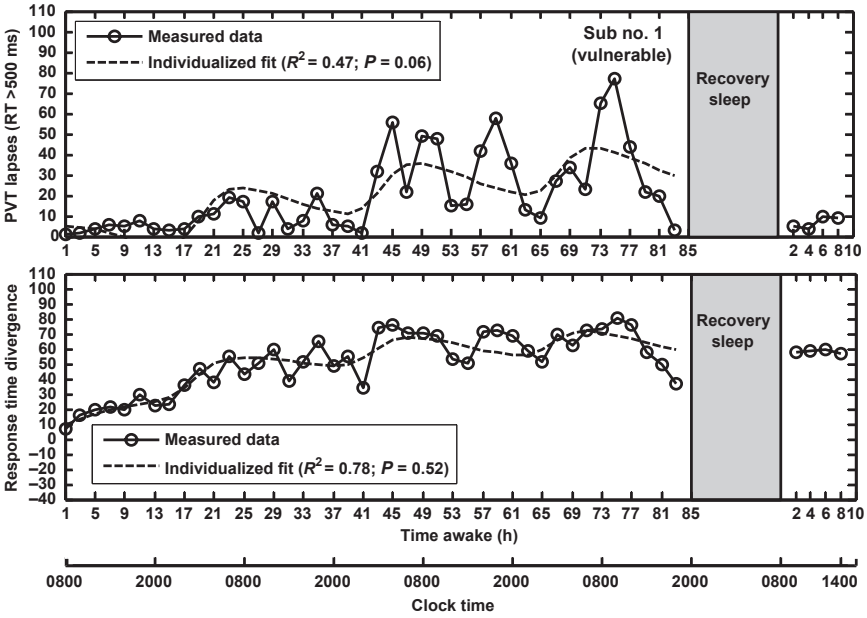

Figure 4. As Fig. 3 for a vulnerable individual (subject 1).

metrics provided statistically similar goodness-of-fit to the two-process model (P = 0.91 versus P = 0.71).

The overall trend for both R^2 and *P*-values indicates that the RTD metric provided a superior goodness-of-fit of the two-process model (i.e. larger R^2 and *P*-values) across all metrics (Table 2). For example, we found that the *P*-values for the RTD metric were larger than those for PVT lapses (RT >500 ms), mean RT, median RT and mean speed for nine, 11, seven and seven subjects, respectively. When we performed pairwise comparisons between the RTD and each

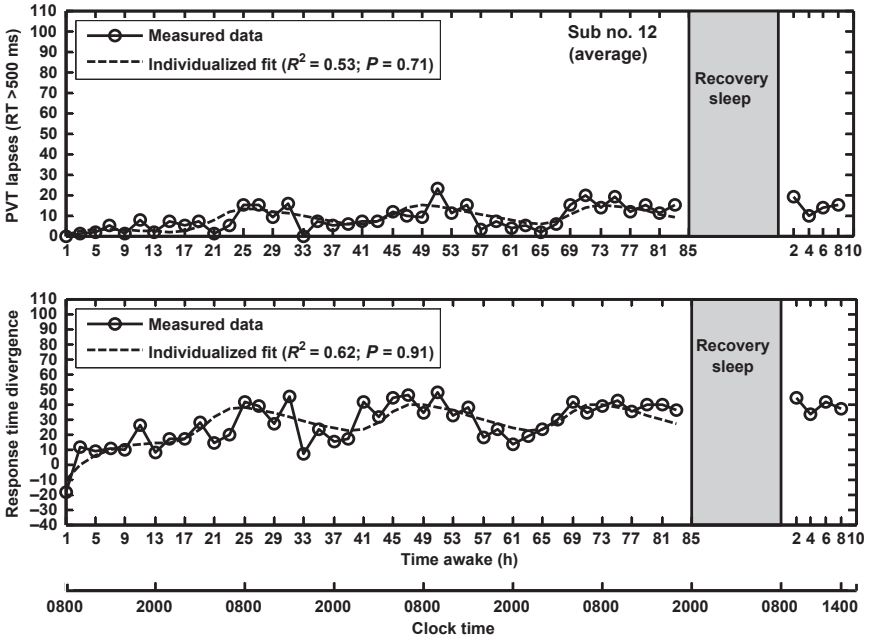

Figure 5. As Fig. 3 for an average individual (subject 12).

of the four other metrics based on R^2 across the 12 subjects using the paired-sample *t*-test (Zar, 1999), we found that the RTD results were statistically larger ($P < 0.05$) than those of all other metrics.

Previously, Wesensten *et al.* (2005) and Lamond *et al.* (2007) analyzed performance restoration after recovery sleep at a group-average level. In this study, we analyzed the performance restoration after 12 h of recovery sleep following 85 h of continuous wakefulness at an individual level and found that recovery is individual-specific and possibly linked to sleep-loss phenotype (Appendix, Section VI).

DISCUSSION

In this paper, we present a new metric (the RTD) for quantifying performance impairment in PVTs due to total sleep deprivation. The RTD has two important advantages over conventionally used PVT metrics. First, it provides the ability to uniquely capture and quantify changes in the entire RT density as well as any specific portion of the density of each PVT session across wakefulness (Fig. 2). In contrast, PVT lapses can reveal changes only in the proportion of responses that fall above a pre-specified threshold but not in

Table 2 Coefficient of determination (R^2) for the individualized two-process model fits for each of the 12 subjects using the five metrics. The *P*-values are for Bartels's test of randomness of the residual errors in the individualized model fits. Larger R^2 and *P*-values indicate better goodness-of-fit to the two-process model of sleep regulation

		R^2 (P-value)				
Subject	*Sleep-loss phenotype*	*RTD*	*PVT lapses (500 ms)*	*Mean RT (ms)*	*Median RT (ms)*	*Mean speed (s^{-1})*
1	Vulnerable	0.78 (0.52)	0.47 (0.06)	0.29 (0.01)	0.44 (0.01)	0.67 (0.08)
2	Vulnerable	0.71 (0.81)	0.40 (0.02)	0.43 (0.01)	0.59 (0.67)	0.65 (0.85)
3	Resilient	0.60 (0.71)	0.33 (0.40)	0.27 (0.12)	0.42 (0.93)	0.51 (0.65)
4	Average	0.44 (0.13)	0.37 (0.46)	0.41 (0.21)	0.44 (0.36)	0.42 (0.20)
5	Average	0.37 (0.58)	0.32 (0.09)	0.35 (0.10)	0.31 (0.68)	0.35 (0.27)
6	Average	0.59 (0.05)	0.59 (0.00)	0.32 (0.00)	0.59 (0.03)	0.60 (0.05)
7	Average	0.80 (0.15)	0.74 (0.33)	0.45 (0.00)	0.38 (0.15)	0.70 (0.04)
8	Average	0.69 (0.69)	0.48 (0.09)	0.34 (0.03)	0.33 (0.09)	0.60 (0.54)
9	Resilient	0.37 (0.12)	0.34 (0.01)	0.25 (0.01)	0.33 (0.03)	0.35 (0.04)
10	Average	0.69 (0.43)	0.60 (0.99)	0.39 (0.01)	0.31 (0.14)	0.58 (0.54)
11	Vulnerable	0.81 (0.04)	0.72 (0.02)	0.59 (0.00)	0.62 (0.05)	0.82 (0.06)
12	Average	0.62 (0.91)	0.53 (0.71)	0.58 (0.79)	0.51 (0.30)	0.59 (0.83)

PVT, psychomotor vigilance test; RT, response time; RTD, response time divergence.

their density. Secondly, unlike existing PVT metrics, the RTD inherently accounts for deviations from a baseline performance level and, therefore, has the potential to capture more effectively the 'true' intersubject variability in PVT performance (Appendix, Section IV). This feature of the RTD metric becomes more significant for sleep-deprivation studies where the subjects' baseline performance levels are expected to vary. In particular, the RTD metric is expected to reflect intersubject variability more clearly than the existing metrics for study groups involving a wide age range, because it has been observed by Philip *et al.* (2004) that, under well-rested conditions (i.e. baseline), older adult subjects respond significantly slower than younger subjects on PVTs.

Using PVT data from a total sleep deprivation laboratory study (Wesensten *et al.*, 2005), we found that PVT lapses over- and underemphasized performance variability systematically and selectively at different times across wakefulness, and that this non-uniformity in performance variability, quantified by the Pearson's correlation coefficient r between the group mean and SD, was dependent on the selected lapse threshold, with r ranging from −0.89 to 0.96 (Fig. 1a–c). We found that raising the threshold resulted in overaccentuating individual performance differences near the trough of the circadian rhythm of alertness (∼04:00–08:00 h), whereas lowering the threshold resulted in overaccentuating individual performance differences near the peak of the circadian rhythm of alertness (∼16:00–20:00 h). This phenomenon can be explained by noting that near the circadian trough the RT densities across individuals vary most in their right tail, whereas near the circadian peak the intersubject variation of RT density is most evident around their central portion (results not shown). Consequently, raising the threshold affects the right tail of the density, accentuating intersubject variability around the circadian trough, whereas lowering the threshold accentuates intersubject variability around the circadian peak. This suggests that the overall performance variability is modulated almost entirely by the selected PVT threshold and raises questions about the selection of the appropriate threshold to quantify performance impairment for a given population. We found the mean and median RT to overemphasize the intersubject variability selectively around the circadian trough and the mean speed to yield relatively uniform variability at all times of day.

In contrast, the RTD metric obviated the need to select a threshold and represented individual differences without over- or underemphasizing them at different times across wakefulness (Fig. 1d). Rather, performance variability appeared constant across all times of day ($r = -0.32$). We attribute this behavior of the RTD metric *vis-à-vis* PVT lapses to the fact that the RTD estimates performance relative to an individual's baseline, thus attenuating the large swings in intersubject variability across the different phases of the circadian rhythm observed in PVT lapses (Appendix, Section IV and Fig. A1). Even when computing the RTD for a portion of the RT density above a specific threshold we found that, unlike PVT lapses, the RTD yielded a more uniform variability across each day, which was arguably independent of time of day and threshold level (Appendix, Section V).

The RTD metric scored the largest effect sizes in capturing the homeostatic and circadian processes underlying performance changes due to sleep loss, followed by mean speed, PVT lapses, median RT and mean RT (Table 1). In general, the effect sizes of mean speed were slightly smaller than those of the RTD metric, while the scores of the other metrics were consistently and considerably smaller than those of the RTD. Further investigation suggests that effect size analysis favors the mean speed metric because of the non-linear, reciprocal transformation of RTs involved in computing this metric. This transformation necessarily deemphasizes the right tail of the RT density—a major source of response variability during prolonged wakefulness among different sleep-loss phenotypes—and, as a result, reduces intersubject variability in mean speed values for each PVT session during wakefulness. Consequently, effect size scores of the mean speed metric increases (Appendix, Section VII).

We assessed indirectly the ability of the metrics to capture the homeostatic and circadian processes by using each metric separately as a dependent variable to fit individual data to the two-process model, which is based on electroencephalography data (Borbély, 1982), and has been validated extensively on PVT data (Rajaraman *et al.*, 2008, 2009; Van Dongen *et al.*, 2007). Again, we found that, in general, the RTD metric yielded consistently better fits than the other four PVT metrics, both in terms of the coefficient of determination (R^2) and the whiteness of the residual errors (Table 2). Because Borbély's two-process model of sleep regulation was not built for a particular performance metric, it should not favor any of the metrics considered in this paper. Hence, to the extent that the two-process model of sleep regulation depicts performance impairment accurately across total sleep deprivation, this indirect metric comparison suggests that the RTD reflects more clearly the interaction

Table 3 Average (n = 12 subjects) Pearson's correlation coefficient between the temporal profiles of each two of the following psychomotor vigilance test (PVT) metrics: response time divergence (RTD), PVT lapses (500 ms), mean response time (RT), median RT and mean speed

Metric	*RTD*	*PVT lapses (500 ms)*	*Mean RT (ms)*	*Median RT (ms)*	*Mean speed (s^{-1})*
RTD	1.00	0.84	0.73	0.85	0.94
PVT lapses (500 ms)	0.84	1.00	0.87	0.90	0.94
Mean RT (ms)	0.73	0.87	1.00	0.88	0.87
Median RT (ms)	0.85	0.90	0.88	1.00	0.94
Mean speed (s^{-1})	0.94	0.94	0.87	0.94	1.00

between homeostatic and circadian processes than the other four PVT metrics.

Because all PVT metrics are derived fundamentally from RT measurements, we analyzed the extent to which the four existing metrics are correlated with each other as well as their correlation with the RTD metric on the basis of the presented study. For the five metrics we considered in this paper, we calculated an average Pearson's correlation coefficient (r) between each pair of metrics as follows. For each subject k, we calculated the values of each metric for each of the 42 trials, and then computed the subject's Pearson's correlation coefficient (r_k) over those 42 data points between all metric pairs. We then took the mean of the values of r_k over the 12 subjects to determine r. Overall, the correlations between metrics range from 0.73 to 0.94. Table 3 reports the value of r for each pair of metrics and indicates that mean speed is correlated highly with each of the other four metrics, and that our proposed RTD metric is correlated less with the other metrics than the other metrics are with each other. We hypothesize that the high correlation of mean speed with all of the other metrics is a result of the non-linear reciprocal transformation used to determine mean speed's value. In our PVT data, high variability in a trial manifests itself as a 'long right tail' of slow responses. When the reciprocal transformation is applied to this long tail, the size of the variance is diminished greatly because a variable that can grow arbitrarily large (i.e. RT) is converted into a variable (i.e. speed) that is bounded both below by zero and above by a finite positive number, as RTs <100 ms are, typically, considered as anticipations and are not included in PVT data analysis. As a result of this non-linear reduction in variance, the spread of the data points across wakefulness is greatly decreased, resulting in a larger value of Pearson's correlation coefficient. The lesser correlation between the RTD and the other metrics is a result of RTD's construction as a whole-distribution metric. The three other metrics capture only a portion of the entire distribution: the PVT lapse metric captures only the right tail of the distribution, while mean RT and median RT are measures of central tendency, not of the entire distribution. We believe that the relatively low correlation between RTD and these metrics is a favorable characteristic of the metric, as it indicates that by calculating a metric using the entire distribution we are capturing not only the aspects captured by PVT lapses, mean RT and median RT, but also other features of the distribution that might be lost when those metrics are applied.

We hypothesize that the new metric may also be indicative of activity in the 'default mode' brain network (Raichle *et al.*, 2001), which has been shown through functional magnetic resonance imaging studies to be associated with the occurrence of slow RTs in PVTs (Drummond *et al.*, 2005) under both well-rested and total sleep-deprived conditions. This is because, under total sleep deprivation, the entire RT distribution changes due mainly to changes in the number and magnitude of slow RTs and due minimally to changes in the number and magnitude of fast RTs (Doran *et al.*, 2001), effectively inducing changes in the right tail of the RT distribution. Accordingly, these changes affect the RTD metric to almost the same extent as the metrics that particularly characterize lapses.

The literature is inconclusive as to what extent sleep-deprived individuals can restore their performance after recovery sleep. Lamond *et al.* (2007) suggest that the conclusions are metric-dependent. Separately, Rosenthal *et al.* (1991) concluded that, after an enforced 24-h time in bed following a 48-h total sleep deprivation routine, subjects could recover only 42% of their total amount of sleep lost, as inferred from polysomnographic recordings. Our own analyses support both possibilities and yield another plausible hypothesis that recovery could be dependent upon the individual's sleep-loss phenotype. Any such analysis, however, is limited by the large variability in PVT results and the lack of statistical power afforded by the small number of observations following recovery sleep. Based on the RTD metric, as well as the mean RT, median RT and mean speed, the results were mixed, suggesting that recovery is individual-specific and possibly linked to sleep-loss phenotype. While all resilient phenotypes recovered, only some of the vulnerable and average ones recovered.

The proposed PVT metric also has some potential limitations. First, to some extent, the utility of the RTD metric depends on the quality and accuracy of the RT density estimates. Although the kernel density technique yielded satisfactory estimates of the RT PDFs, this and other techniques may fail to obtain accurate estimates when the RTs observed in a session are few in number. This may occur in short PVT sessions, e.g. <10 min, or with increased sleepiness, as the total number of stimuli presented to a subject in a session decreases with increasing RTs. Secondly, there is no consensus definition of what constitutes baseline data. Nevertheless, we found the proposed metric to be only mildly sensitive to the selection of baseline data. For example, when we defined baseline density alternatively as the density of the aggregate RT values of four sessions (14:00, 16:00, 18:00 and 20:00 h) on day 1 and recomputed the RTD for each of the 46 (42 plus four) PVT sessions, we found that in 94% (43 of 46) of the sessions the results were statistically indistinguishable (based on paired-sample t-test, $P < 0.05$) from those obtained when we selected baseline densities as those which were as close as possible to Gaussian. This suggests that reasonable alternative definitions of what constitute baseline data should provide similar results. Another potential limitation of the RTD metric is its complexity. In an attempt to alleviate such limitation, we made available the set of MATLAB computer programs used for computing the proposed metric, which can be downloaded from http://bhsai.org/bic/papers/RTD_suite.zip.

In summary, we conclude that the RTD metric provides a number of advantages over the traditional PVT metrics for quantifying PVT performance impairment due to sleep loss: the RTD does not pose a quandary in the selection of an appropriate threshold, it does not over- or underemphasize

performance variability at different times of day and it captures more accurately the interaction between the sleep homeostatic and circadian processes. We also conclude that as the existing PVT metrics lack the ability to reflect performance deviations relative to a baseline level, they may not properly account for intersubject variability in performance impairment. In addition, this work raises an intriguing question: for the threshold-based metrics, is the observed dependency between the threshold value and the corresponding correlation between performance variability and the circadian phase a true characteristic of a subject's cognitive state, or is it simply an idiosyncrasy of the PVT threshold metric itself? We argue that the answer to this question should not be metric-dependent and that the quantitative analyses reported here provide evidence for a more careful interpretation of PVT data and their implications on performance impairment.

ACKNOWLEDGEMENTS

This work was funded, in part, by the Military Operational Medicine Research Area Directorate of the US Army Medical Research and Materiel Command, Fort Detrick, MD.

DISCLAIMER

The opinions and assertions contained herein are the personal views of the authors and are not to be construed as official or as reflecting the views of the US Army or of the US Department of Defense. This paper has been approved for public release with unlimited distribution. This was not an industry-supported study. All authors have reported no financial conflicts of interest.

REFERENCES

Bartels, R. The rank version of von Neumann's ratio test for randomness. *J. Am. Stat. Asscoc.*, 1982, 77: 40–46.

Basner, M. and Dinges, D. F. Maximizing sensitivity of the psychomotor vigilance test (PVT) to sleep loss. *Sleep*, 2011, 34: 581–591.

Borbély, A. A. A two process model of sleep regulation. *Hum. Neurobiol.*, 1982, 1: 195–204.

Box, G. E. P. and Cox, D. R. An analysis of transformations. *J. Roy. Stat. Soc. B. Met.*, 1964, 26: 211–252.

Chatfield, C. *The Analysis of Time Series: An Introduction*, 6th edition. Chapman & Hall/CRC, Boca Raton, 2004.

Dinges, D. F. and Powell, J. W. Microcomputer analyses of performance on a portable, simple visual RT task during sustained operations. *Behav. Res. Meth. Ins. C.*, 1985, 17: 652–655.

Doran, S. M., Van Dongen, H. P. A. and Dinges, D. F. Sustained attention performance during sleep deprivation: Evidence of state instability. *Archives Italiennes De Biologie*, 2001, 139: 253–267.

Dorrian, J., Rogers, N. L. and Dinges, D. F. Psychomotor vigilance performance: neurocognitive assay sensitive to sleep loss. In: C. A. Kushida (Ed.) *Sleep Deprivation. Clinical Issues, Pharmacology, and Sleep Loss Effects*. Marcel Dekker, New York, 2005: 39–70.

Drummond, S., Bischoff-Grethe, A., Dinges, D., Ayalon, L., Mednick, S. and Meloy, M. The neural basis of the psychomotor vigilance task. *Sleep*, 2005, 28: 1059–1068.

Efron, B. and Tibshirani, R. *An Introduction to the Bootstrap*. Chapman & Hall, New York, 1993.

Endres, D. M. and Schindelin, J. E. A new metric for probability distributions. *IEEE T. Inform. Theory*, 2003, 49: 1858–1860.

Hedges, L. V. and Olkin, I. *Statistical Methods for Meta-analysis*. Academic Press, Orlando, 1985.

Holden, J. G., Van Orden, G. C. and Turvey, M. T. Dispersion of response times cognitive dynamics. *Psychological Review*, 2009, 116: 318–342.

Kullback, S. and Leibler, R. A. On information and sufficiency. *Ann. Math. Stat.*, 1951, 22: 79–86.

Lamond, N., Jay, S. M., Dorrian, J., Ferguson, S. A., Jones, C. and Dawson, D. The dynamics of neurobehavioural recovery following sleep loss. *J. Sleep Res.*, 2007, 16: 33–41.

Lin, J. Divergence measures based on the Shannon entropy. *IEEE T. Inform. Theory*, 1991, 37: 145–151.

Morris, S. B. and DeShon, R. P. Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychol. Meth.*, 2002, 7: 105–125.

Philip, P., Taillard, J., Sagaspe, P. *et al.* Age, performance and sleep deprivation. *J. Sleep Res.*, 2004, 13: 105–110.

Raichle, M. E., Macleaod, A. M., Synder, A. Z., Powers, W. J., Gusnard, D. A. and Shulman, G. L. A default mode of brain function. *Proc. Natl Acad. Sci. USA*, 2001, 98: 676–682.

Rajaraman, S., Gribok, A. V., Wesensten, N. J., Balkin, T. J. and Reifman, J. Individualized performance prediction of sleep-deprived individuals with the two-process model. *J. Appl. Physiol.*, 2008, 104: 459–468.

Rajaraman, S., Gribok, A. V., Wesensten, N. J., Balkin, T. J. and Reifman, J. An improved methodology for individualized performance prediction of sleep-deprived individuals with the two-process model. *Sleep*, 2009, 32: 1377–1392.

Rosenthal, L., Merlotti, L., Roehrs, T. A. and Roth, T. Enforced 24-hour recovery following sleep deprivation. *Sleep*, 1991, 14: 448–453.

Silverman, B. W. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London/New York, 1986.

Van Dongen, H. P. A., Mott, C. G., Huang, J. K., Mollicone, D. J., McKenzie, F. D. and Dinges, D. F. Optimization of biomathematical model predictions of cognitive performance impairment in individuals: accounting for unknown traits and uncertain states in homeostatic and circadian processes. *Sleep*, 2007, 27: 423–433.

Van Zandt, T. How to fit a response time distribution. *Psychon. B. Rev.*, 2000, 7: 424–465.

Wesensten, N. J., Killgore, W. D. S. and Balkin, T. J. Performance and alertness effects of caffeine, dextroamphetamine, and modafinil during sleep deprivation. *J. Sleep Res.*, 2005, 14: 255–266.

Zar, J. H. *Biostatistical Analysis*, 4th edition. Prentice Hall, Upper Saddle River, NJ, 1999.

APPENDIX

Section I. Empirical estimation of response time (RT) probability density function (PDF)

Among the most widely used techniques to estimate PDFs empirically are those that are based on kernel density estimation (Silverman, 1986) in which a kernel (or weighting) function, usually represented by a standard Gaussian density and characterized by a smoothing parameter, is used to approximate the density around each observation in the sample. Mathematically, the density function estimate $\hat{g}(t)$ of a series of *N* independent and identically distributed RT samples $\{t_n\}_1^N$ observed in a psychomotor vigilance test (PVT) session can be expressed as follows:

$$\hat{g}(t) = \frac{1}{Nh}\sum_{n=1}^{N} K\left(\frac{t - t_n}{h}\right), \qquad (A1)$$

where h is a positive, real number smoothing parameter and $K(\cdot)$ is a kernel function. Here, we used a standard Gaussian density for $K(\cdot)$. In this formulation, as h approaches zero, the density estimates exhibit larger values around the observations t_n, whereas as h becomes larger the density estimates become smoother, approaching a flat, uniform density function.

Naturally, for a long-tailed (especially the right tail) RT density, such as those observed during sleep deprivation, the density estimates from eqn (A1) result in either undersmoothed tails or oversmoothed central portions of the RT density (Silverman, 1986). To address this problem, we used an alternate technique for inferring $\hat{g}(t)$ in which we first transformed the RT data through the Box–Cox family of transformations (Box and Cox, 1964). This technique attempts to map monotonically the original (measured) RT data set $\{t_n\}_1^N$, for each PVT session of each subject, into a transformed, normally distributed (to the maximum extent possible) data set $\{t'_n\}_1^N$. Assuming that $\{t'_n\}_1^N$ was sufficiently distributed normally (at least the skewness and excess kurtosis were reduced), we used eqn (A1) to compute the corresponding empirical density estimate $\hat{g}'(t')$, where the smoothing parameter h was computed separately for each PVT session of each subject using eqn (A1) (Silverman, 1986). We then computed $\hat{g}(t)$ by converting $\hat{g}'(t')$ through the corresponding inverse transformation. While the proposed method allows for the estimation of the RT density using the entire set of RTs obtained during a PVT session, it can be modified readily to estimate the density of any portion of the RT data.

Section II. Quantification of the dissimilarity between two RT PDFs

We used the Jensen–Shannon divergence (JSD; Lin, 1991), which is derived from the Kullback–Leibler divergence (KLD; Kullback and Leibler, 1951), to quantify the dissimilarity between any two RT PDFs, e.g. $p_1(t)$ and $p_2(t)$, where t represents a random variable (i.e. the RT). The JSD is defined mathematically as follows:

$$\mathrm{JSD}\left[p_1(t), p_2(t)\right] = \frac{1}{2}\mathrm{KLD}\left(p_1(t), \frac{[p_1(t) + p_2(t)]}{2}\right) + \frac{1}{2}\mathrm{KLD}\left(p_2(t), \frac{[p_2(t) + p_1(t)]}{2}\right). \qquad (A2)$$

In this formulation, as $p_1(t)$ converges to $p_2(t)$ or *vice versa*, the JSD$[p_1(t), p_2(t)]$ approaches zero. Conversely, as $p_1(t)$ diverges from $p_2(t)$, the JSD$[p_1(t), p_2(t)]$ increases, attaining a maximum value of ln(2), where ln(·) is the natural logarithm. The $\sqrt{\mathrm{JSD}[p_1(t), p_2(t)]}$ has been established recently as a formal metric for quantifying the distance between two PDFs because it possesses all the necessary attributes of a robust dissimilarity metric, which are missing in the KLD (Endres and Schindelin, 2003).

To compute the dissimilarity between two PDFs for a desired portion of the observed responses (akin to PVT lapses), eqn (A2) can be rewritten in a more general form as follows:

$$\underset{[a-b]}{\mathrm{JSD}}[p_1(t), p_2(t)] = \frac{1}{2}\mathrm{KLD}\left(p_1(t)/k_1, \frac{[p_1(t)/k_1 + p_2(t)/k_2]}{2}\right) + \frac{1}{2}\mathrm{KLD}\left(p_2(t)/k_2, \frac{[p_2(t)/k_2 + p_1(t)/k_1]}{2}\right), \qquad (A3)$$

where [a–b], with $0 \le a < b < \infty$, represents the interval containing the portion of responses we want to compare between the two PDFs, $k_1 = \int_a^b p_1(t)\,dt$ and $k_2 = \int_a^b p_2(t)\,dt$. Note that as a approaches 0 and b approaches ∞, both k_1 and k_2 converge to 1 and eqn (A3) converges to eqn (A2).

As any distance metric, the JSD quantifies only the dissimilarity between the entire set of RTs observed during any two PVT sessions but does not convey whether performance in one session is better or worse than that in the other session. To account for the direction of the change in performance from one session to the other, we analyzed whether the RTs in one PVT session were statistically greater than, equal to or less than the RTs in the other session and, accordingly, determined whether performance deteriorated or improved from session to session. The expression for the RTD for the portion of responses within the interval [a–b] at PVT session T is as follows:

$$\mathrm{RTD}(T) = \mathrm{sgn}[p_T(t), p_{BL}(t)]\frac{N_T}{\sqrt{\ln(2)}}\sqrt{\underset{[a-b]}{\mathrm{JSD}}\,[p_T(t), p_{BL}(t)]}, \qquad (A4)$$

where N_T represents the total number of responses observed in PVT session T, $p_T(t)$ and $p_{BL}(t)$ represent the RT PDFs at session T and the baseline session, respectively and sgn$[p_T(t), p_{BL}(t)]$ represents a function that takes the value of 1.0 when the RTs in PVT session T are statistically greater than those in the baseline session and −1.0 otherwise. This allowed us to identify whether performance at T was worse {i.e. sgn$[p_T(t), p_{BL}(t)] = 1.0$} or as good as or better than {sgn$[p_T(t), p_{BL}(t)] = -1.0$} the baseline performance level, where we used the two-sample Kolmogorov–Smirnov (KS) test (Zar, 1999) to determine the value of sgn$[p_T(t), p_{BL}(t)]$. The proportionality constant $N_T/\sqrt{\ln(2)}$ in eqn (A4) was used so that the magnitude of the proposed RTD metric varied within the same range as PVT lapses, i.e. from zero to N_T, facilitating direct comparisons between the two metrics. Based on this definition, the proposed PVT metric would

have negative values when performance is better than or equivalent to the baseline level and positive values when performance is worse than the baseline level.

Section III. Measures for comparing and contrasting PVT performance metrics

In this section, we describe the measures used for comparing and contrasting the PVT performance metrics in their ability to capture the homeostatic and circadian processes underlying sleep regulation.

Effect size

To assess the effect size of sleep deprivation on performance within subjects (or repeated measures) for a group of individuals, we considered the difference in performance between well-rested and sleep-deprived conditions for each individual first, and then averaged these pairwise differences across the group (Morris and DeShon, 2002). Here, we followed the procedure proposed by Basner and Dinges (2011), using the expressions defined in Morris and DeShon (2002) to compute the effect size d_{bp} for a group of individuals between performance at sessions b (at the baseline) and p (post-baseline), so that:

$$d_{bp} = \frac{\bar{y}_p - \bar{y}_b}{\sqrt{\frac{1}{M}\sum_{m=1}^{M}\left[y_{pm} - y_{bm} - (\bar{y}_p - \bar{y}_b)\right]^2}}, \qquad (A5)$$

where y_{pm} and y_{bm} represent the PVT performance measurements of the mth individual at the post-baseline and baseline sessions, respectively, and $\bar{y}_p$ and $\bar{y}_b$ denote the mean PVT performance computed over the M individuals at the post-baseline and baseline sessions, respectively. To compute the overall (average) effect size over P PVT sessions, we used the following expression (Hedges and Olkin, 1985):

$$\bar{d} = \frac{\sum_{p=1}^{P} d_{bp} w_{bp}}{\sum_{p=1}^{P} w_{bp}}, \qquad (A6)$$

where w_{bp} represents the inverse of the sampling variance of d_{bp}, which was estimated by bootstrapping (Efron and Tibshirani, 1993).

Goodness of two-process model fit

To assess the goodness-of fit-of each metric to Borbély's two-process model output (Borbély, 1982), we used two goodness-of-fit measures. The first was the coefficient of determination R^2, which quantifies the proportion of the variance in an individual's performance data explained by a model (Zar, 1999); in this case, the individualized two-process model fit of performance impairment (Rajaraman *et al*., 2008, 2009). For the mth individual across J PVT sessions, this measure was expressed mathematically as follows:

$$R^2 = 1 - \frac{\sum_{j=1}^{J}(y_{jm} - P_{jm})^2}{\sum_{j=1}^{J}(y_{jm} - \bar{y}_m)^2}, \qquad (A7)$$

where y_{jm} and P_{jm} represent the PVT performance measurement and corresponding two-process model fit, respectively, for the jth PVT session of the mth individual, and $\bar{y}_m$ represents the mth individual's mean performance across J PVT sessions administered during wakefulness. In this formulation, R^2 ranges from negative ∞ to 1.0, and the higher the R^2 value, the better the data fit the model.

The second goodness-of-fit measure was the degree of whiteness (randomness) of the residual error, i.e. the difference between the individualized two-process model fit and the performance measurements, where the more random the residual error, the more accurate the model describes the data, regardless of its scale (Chatfield, 2004). Here, we used Bartels's rank test (Bartels, 1982; $P < 0.05$) to determine the degree of randomness of the residual error, where we tested the null hypothesis that the residual error was a random signal versus the alternative hypothesis that it was a non-random signal.

Section IV. The effect of considering baseline in quantifying PVT performance

We hypothesize that by estimating performance relative to an individual's baseline, the RTD attenuates the large swings in intersubject variability across different phases of circadian rhythm in PVT lapses. To test this hypothesis, we analyzed the intersubject variability between two subjects with 'resilient' and 'vulnerable' sleep-loss phenotypes in two scenarios: (1) during the early evening hours, where PVT lapses (RT >500 ms) suppress intersubject variability and (2) during the early morning hours, where PVT lapses accentuate intersubject variability. Figs 3 and 4 (main text) show the PVT lapses for RT >500 ms (open circles, top) and the RTD metric (open circles, bottom) for the two subjects with resilient (subject 9; Table 2, main text) and vulnerable (subject 1; Table 2) sleep-loss phenotypes, respectively. We estimated an individual's sleep-loss phenotype as the sum of the individual's homeostatic component [$\alpha S(t)$ (in Rajaraman *et al*., 2008, 2009)] at 82 h of wakefulness and the circadian amplitude (β), both estimated from the two-process model fit (Rajaraman *et al*., 2008, 2009) to the individual's RTD data. We categorized the individuals with values within the mean ± standard deviation (SD) as average sleep-loss phenotypes, and those above and below this range as vulnerable and resilient, respectively (Table 2).

Figs 3 and 4 show that at 37 h of wakefulness (early evening hour, 20:00 h, day 2) the lapses for the two subjects were nearly identical (four for subject 9 and six for subject 1), whereas the RTD values for these subjects were significantly different (17 for subject 9 and 49 for subject 1). To investigate the reason behind this discrepancy, we analyzed the RT densities for the two subjects at 37 h of wakefulness. Fig. A1 (top) shows the empirical RT density estimates at 37 h of wakefulness for subject 9 (thin solid blue line) and subject 1 (thin dashed red line), which appeared to be very close to each other in the region to the right of 500 ms, corroborating the nearly identical results for PVT lapses. Fig. A1 (top) also shows that for subject 9 the dissimilarity between the entire RT PDF at baseline and at 37 h (solid blue lines) was relatively significantly smaller than the corresponding dissimilarity for subject 1 (dashed red lines), illustrating why the RTD values for these subjects were significantly different (17 versus 49) at 37 h. Conversely, at 47 h of wakefulness (early morning hour, 06:00 h, day 3) shown in Figs 3 and 4 (main text), we found that while the difference in lapses between the two subjects was significant (50 for subject 9 and 22 for subject 1), the RTD attenuated this difference (59 for subject 9 and 70 for subject 1). This is illustrated in Fig. A1 (bottom), which shows that the dissimilarity between the RT PDFs at 47 h relative to their baselines was somewhat equivalent for the two subjects, while the differences for RTs >500 ms were more significant. We concluded, therefore, that by estimating performance relative to an individual's baseline, the RTD attenuated the large swings in intersubject variability [differences of 32 (49–17) at 37 h and of 11 (70–59) at 47 h], which were accentuated by PVT lapses [differences of 2 (6–4) at 37 h and of 28 (50–22) at 47 h] at early evening and morning hours.

Section V. Head-to-head comparison between PVT lapses and the RTD metric

To facilitate head-to-head comparisons between PVT lapses (see Fig. 1a–c, main text) and the RTD metric, we computed the RTD for specific portions of the RT density corresponding to the associated thresholds. Fig. A2 shows the mean and SD of the RTD metric results for: (a) RT >1000 ms, (b) RT >500 ms and (c) RT >250 ms, obtained from eqn (A4) by setting *a* to 1000, 500 and 250 ms, respectively, and *b* to 600,000 ms. Fig. A2a shows that the mean RTD results for RT >1000 ms failed to describe the homeostatic build-up and

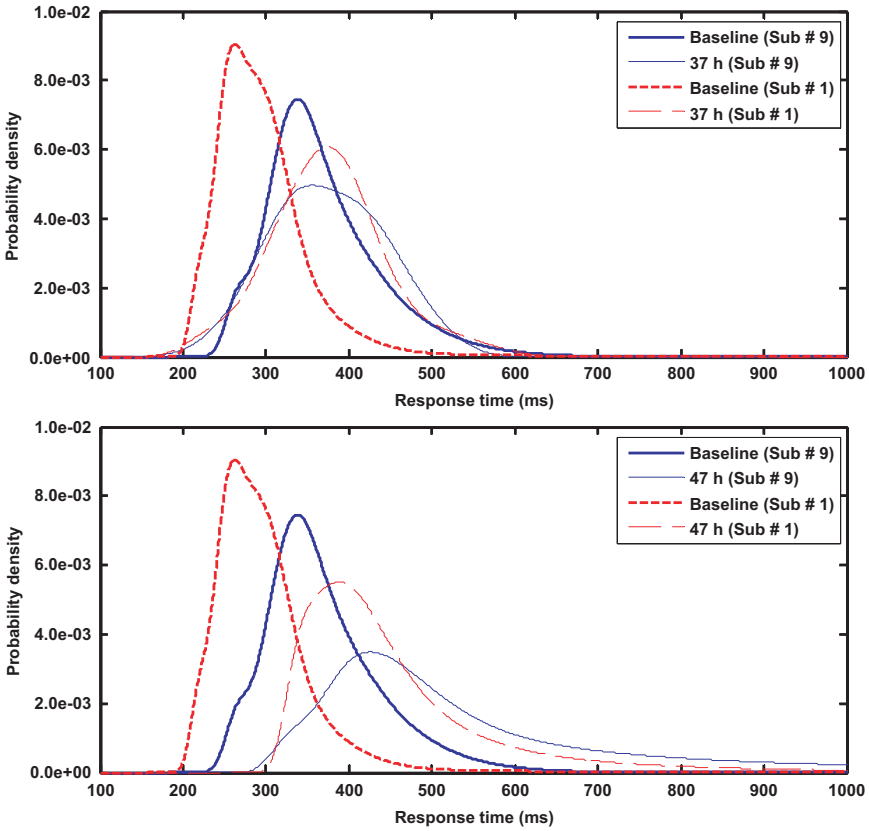

Figure A1. The empirical probability density estimates of the response time (RT) measurements for a resilient subject (no. 9; solid blue lines) and a vulnerable subject (no. 1; dashed red lines) at 37 h (early evening hour; top) and at 47 h (early morning hour; bottom) of wakefulness. At 37 h, subject 9 had lapses (>500 ms) = 4 and response time divergence (RTD) = 17, while subject 1 had lapses = 6 and RTD = 49. The small difference in lapse values of the two subjects is reflected by the overlap of their 37-h densities for RT >500 ms, while the relatively large difference in RTD is reflected by the dissimilarity of the corresponding densities at baseline and 37 h of wakefulness. At 47 h, subject 9 had lapses = 50 and RTD = 59, while subject 1 had lapses = 22 and RTD = 70.

Table A1 Differences in mean performance between averages over sessions at 10:00, 12:00, 14:00 and 16:00 h on day 5 and the corresponding sessions on day 1, for the response time divergence (RTD), psychomotor vigilance test (PVT) lapses (500 ms), mean response time (RT), median RT and mean speed metrics for each of the 12 subjects. (Note that the direction of change of mean speed is opposite to those of the other metrics.) The *P*-values are for the paired-sample *t*-tests, where smaller *P*-values indicate a larger statistical significance of the differences in mean performances

	Difference of the means (day 5 - day 1) (P-value)				
Subject	*RTD*	*PVT lapses (500 ms)*	*Mean RT (ms)*	*Median RT (ms)*	*Mean speed (s^{-1})*
1	38.88 (0.00)	2.75 (0.56)	67.90 (0.00)	70.88 (0.01)	−0.61 (0.00)
2	11.85 (0.48)	−0.50 (0.39)	9.48 (0.41)	15.75 (0.09)	−0.15 (0.19)
3	4.18 (0.81)	1.50 (0.19)	2.96 (0.91)	5.50 (0.81)	−0.05 (0.80)
4	−21.04 (0.31)	−1.25 (0.06)	−25.85 (0.28)	−24.88 (0.10)	0.26 (0.23)
5	−28.00 (0.04)	−1.00 (0.92)	−21.13 (0.02)	−23.13 (0.00)	0.29 (0.01)
6	8.12 (0.42)	1.75 (0.12)	12.00 (0.40)	9.50 (0.57)	−0.09 (0.55)
7	44.60 (0.03)	27.25 (0.06)	151.59 (0.07)	88.50 (0.07)	−0.54 (0.08)
8	−10.76 (0.33)	−0.25 (0.65)	−17.03 (0.42)	−11.00 (0.38)	0.12 (0.44)
9	−43.30 (0.00)	−3.25 (0.01)	−41.43 (0.00)	−32.75 (0.00)	0.30 (0.00)
10	12.17 (0.00)	8.25 (0.06)	46.33 (0.10)	32.75 (0.01)	−0.21 (0.00)
11	−1.93 (0.92)	1.50 (0.39)	13.22 (0.69)	1.63 (0.94)	−0.06 (0.81)
12	28.97 (0.00)	12.25 (0.39)	82.32 (0.00)	65.50 (0.01)	−0.51 (0.01)

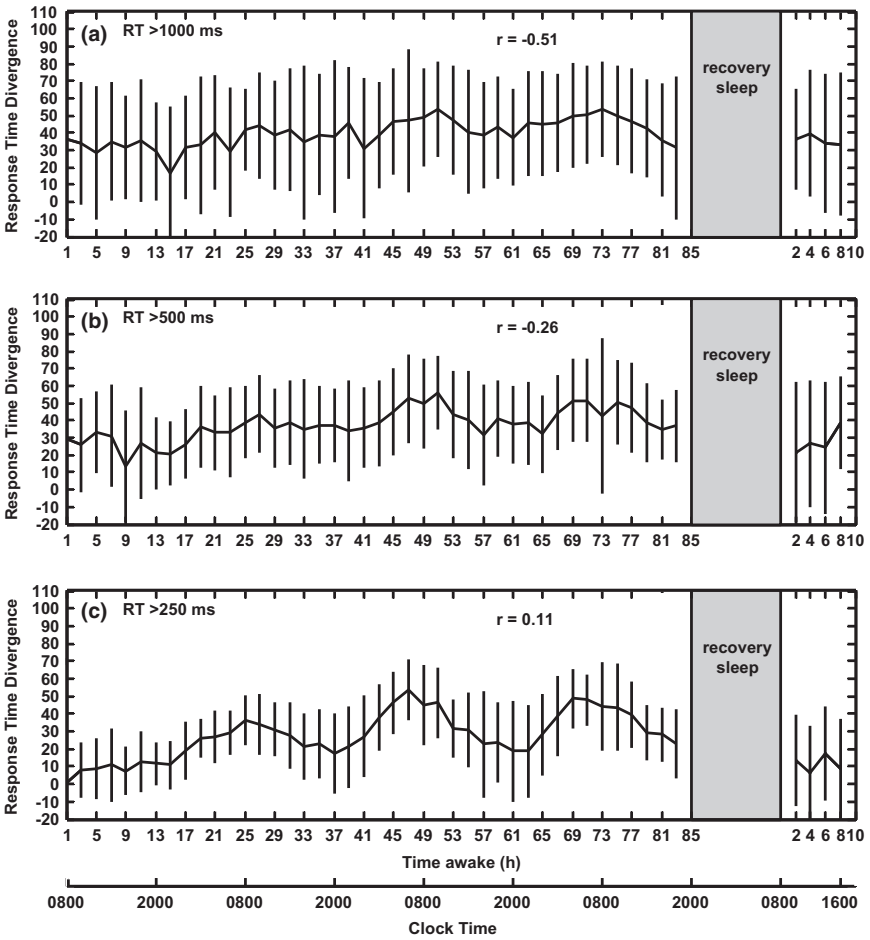

Figure A2. Mean and standard deviation (SD) of the response time divergence (RTD) metric ($n = 12$) computed from the empirical density estimate for three different thresholds: (a) 1000 ms, (b) 500 ms and (c) 250 ms. We used Pearson's correlation coefficient (r) to compute the correlation between the temporal profiles of the mean and SD as a function of wakefulness. The shaded region in each plot represents the 12-h recovery sleep administered to the subjects immediately after 85 h of total sleep deprivation.

circadian rhythm underlying performance impairment. Rather, it varied inconsistently across wakefulness, suggesting that for RT >1000 ms the dissimilarity in RT densities is random (i.e. originating chiefly from intra-individual variability). Moreover, we observed that the SD of the RTD results was approximately constant across wakefulness, with no significant correlation with the mean ($r = -0.51$). Similarly, we found that the mean RTD results for RT >500 ms (Fig. A2b) also varied randomly with wakefulness but to a lesser extent than that for the RT >1000 ms threshold, again with no significant correlation with the corresponding SD ($r = -0.26$). Conversely, we found that the mean RTD results for RT >250 ms (Fig. A2c) captured more effectively the homeostatic and circadian processes, with the corresponding SD varying uniformly across wakefulness and having no significant correlation with the mean ($r = 0.11$). These results suggest that the information about the cognitive state of an individual described by a portion of the density above a threshold increased as we lowered the threshold. This is consistent with the notion that the entire density of RTs should be used to characterize and assess performance decrements due to sleep loss. As we lowered the threshold, we included a larger portion of the RT density in computing the RTD metric, thereby quantifying PVT performance more comprehensively. For example, group RTD results for RT >0 ms (Fig. 1d, main text) were statistically indistinguishable from group RTD results for RT >125 ms (results not shown), as inferred from the paired-sample *t*-test (Zar, 1999; $P < 0.05$) performed across the two group measures for each PVT session. That is, we found no apparent change in the group PVT performance as we lowered the threshold below 125 ms.

Section VI. Performance restoration after recovery sleep

To assess the extent of performance restoration to pre-sleep deprivation levels after recovery sleep for each subject, we tested the significance of the differences in performance at the four sessions after the 12-h recovery sleep, i.e. at 10:00, 12:00, 14:00 and 16:00 h on day 5 and the corresponding sessions on day 1, for all five PVT metrics, using the paired-sample *t*-test (Zar, 1999). The entries in Table A1 show the difference of the means (mean day 5–mean day 1) and their corresponding *P*-values for the five metrics for each of the 12 subjects. (Note that the direction of change of mean speed is opposite to those of the other metrics.) In stark contrast to the other metrics, for PVT lapses, only subject 9 (Table 2, main text) had a significant difference in the means ($P < 0.05$); however, the difference was negative (−3.25; Fig. 3, main text), indicating that performance was restored for all subjects after 12 h of recovery sleep. For the RTD as well as for the mean RT, median RT and mean speed the results were mixed, suggesting that recovery is individual-specific and linked possibly to sleep-loss phenotype. Based on these metrics, between eight and 10 subjects seemed to have recovered, as they had either no significant differences in the means or when the difference in the means was significant ($P < 0.05$) it was negative for RTD, mean RT and median RT and positive for mean speed (e.g. subjects 5 and 9).

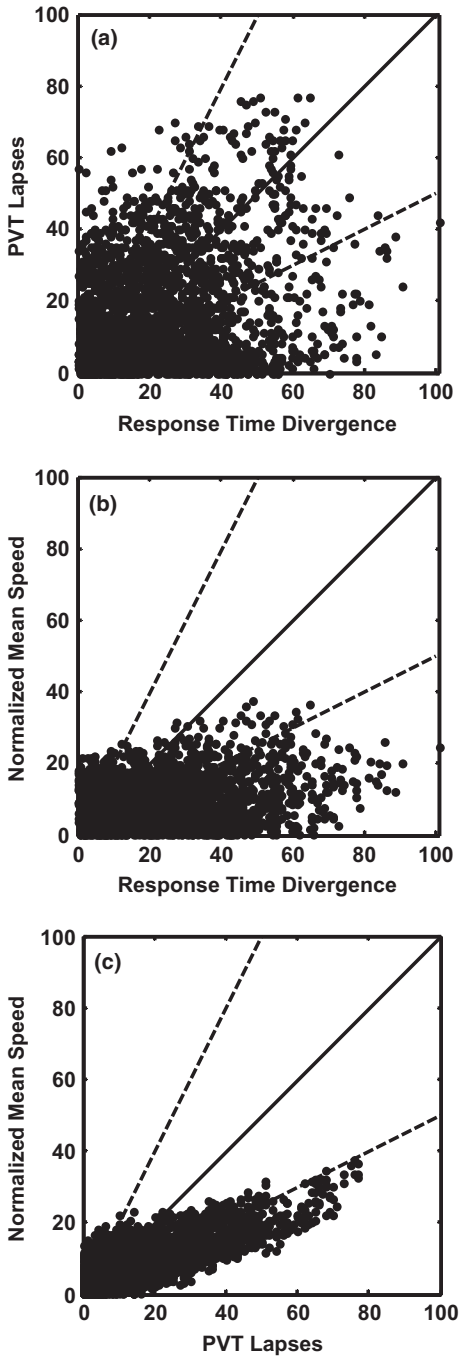

Figure A3. Comparison of the magnitude of the intersubject performance differences between each of the 66 ($12 \times 11 \div 2$) possible distinct pairs of individuals for each of the 42 psychomotor vigilance test (PVT) sessions during wakefulness for: (a) PVT lapses [response time (RT) >500 ms] versus the response time divergence (RTD) metric, (b) normalized mean speed versus the RTD metric and (c) normalized mean speed versus PVT lapses. In each panel, the solid line has a slope = 1, whereas the dashed lines above and below the solid line have slopes equal to 2 and 1/2, respectively.

Reassuringly, all but the PVT lapse metric provide strong statistical support that subjects 1 and 12, with vulnerable and average sleep-loss phenotypes, respectively, did not recover, while the two resilient subjects (no. 3 and no. 9) recovered.

Section VII. How effect size analysis favors the mean speed metric

We found that the effect size analysis favors the mean speed metric. This is possibly because the non-linear, reciprocal transformation involved in computing mean speed necessarily de-emphasizes the right tail of the RT density, which contains the majority of response variability during prolonged wakefulness among sleep-loss phenotypes. This results in the reduction of intersubject variability in mean speed values for each PVT session during wakefulness. Consequently, the variance of the pairwise (i.e. within-subject) differences in mean speed across a group of individuals also decreases, reducing the denominator in eqn (A5) and effectively increasing the effect size scores of the mean speed metric. To investigate further this observation, we performed an additional analysis where, for each metric, we computed the intersubject performance differences between each pair of subjects for each of the 42 PVT sessions. As illustrated in Fig. A3, where we compare the magnitude of the intersubject performance differences between PVT lapses, RTD and normalized mean speed, mean speed yielded a $\sim$50% smaller intersubject performance difference when compared with those of PVT lapses and the RTD metric (Fig. A3b,c). This supports our finding that, by construction, the mean speed metric reduces intersubject variability significantly, thus favoring effect size analysis.