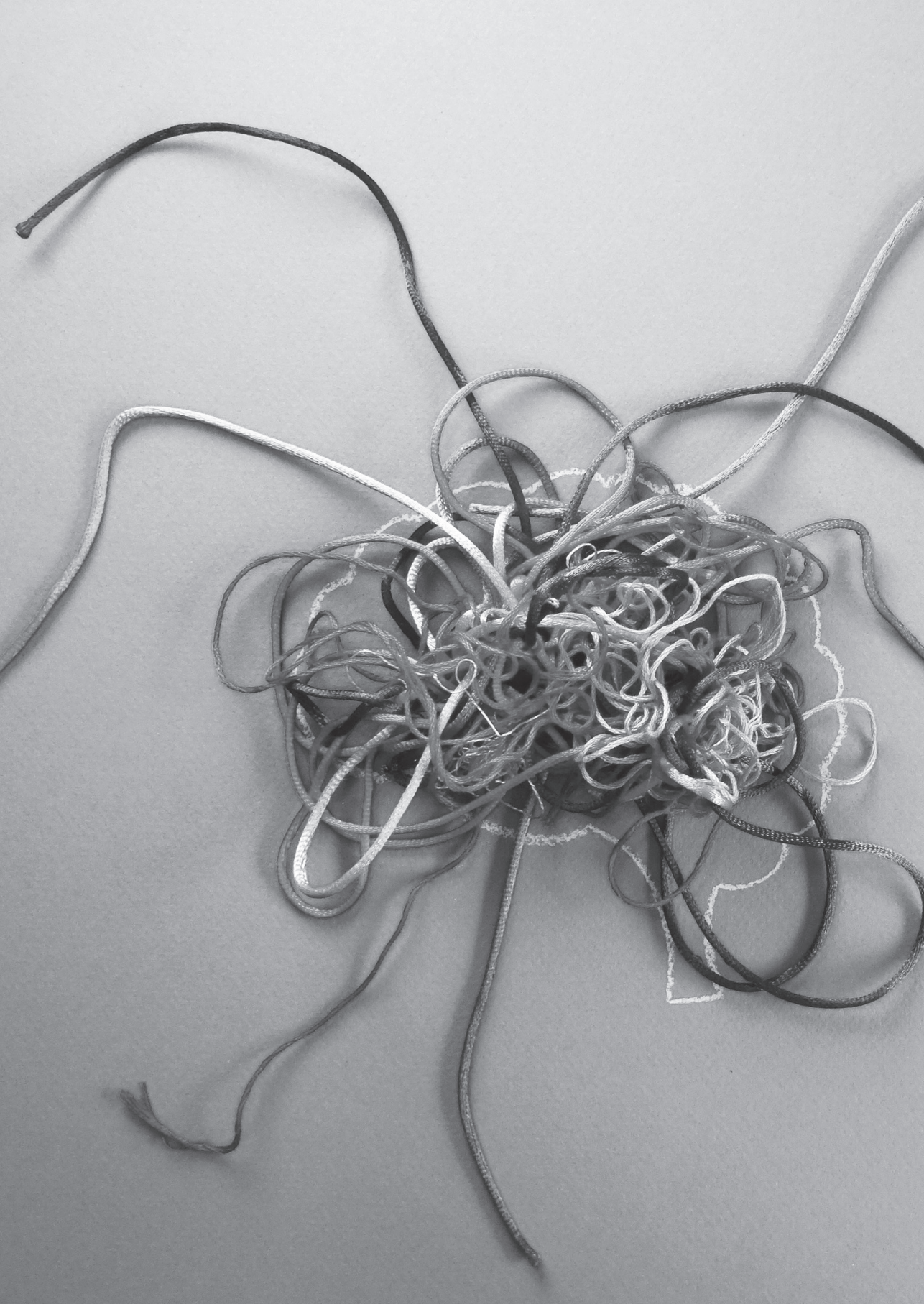
Unravelling narcolepsy : from pathophysiology to measuring treatment effects

Heide, A. van der



CHAPTER 6

COMPARING TREATMENT EFFECT MEASUREMENTS IN NARCOLEPSY: THE SUSTAINED ATTENTION TO RESPONSE TASK, EPWORTH SLEEPINESS SCALE AND MAINTENANCE OF WAKEFULNESS TEST

Astrid van der Heide
Mojca K.M. van Schie
Gert Jan Lammers
Yves Dauvilliers
Isabelle Arnulf
Geert Mayer
Claudio L. Bassetti
Claire-Li Ding
Philippe Lehert
J. Gert van Dijk

Sleep 2015;38(7):1051–8

ABSTRACT

Study objectives: To validate the Sustained Attention to Response Task (SART) as a treatment effect measure in narcolepsy, and to compare the SART with the Maintenance of Wakefulness Test (MWT) and the Epworth Sleepiness Scale (ESS).

Design: Validation of treatment effect measurements within a randomised controlled trial (RCT).

Patients: 95 patients with narcolepsy with or without cataplexy.

Interventions: The RCT comprised a double-blind, parallel-group, multi-centre trial comparing the effects of 8-week treatments with pitolisant (BF2.649), modafinil or placebo (NCT01067222). MWT, ESS and SART were administered at baseline and after an 8-week treatment period. The severity of excessive daytime sleepiness and cataplexy was also assessed using the Clinical Global Impression scale (CGI-C).

Measurements and results: The SART, MWT and ESS all had good reliability, obtained for the SART and MWT using two to three sessions in one day. The ability to distinguish responders from non-responders, classified using the CGI-C score, was high for all measures, with a high performance for the SART (r=0.61) and the ESS (r=0.54).

Conclusions: The SART is a valid and easy to administer measure to assess treatment effects in narcolepsy, enhanced by combining it with the ESS.

INTRODUCTION

While narcolepsy has an undisputed profound impact on daily life,[1] quantifying how it impairs daily life is difficult. The severity of narcolepsy is currently assessed using measures of the ability to stay awake in boring conditions, such as the Maintenance of Wakefulness Test (MWT), or measures of subjective sleepiness, for which the Epworth Sleepiness Scale (ESS) is often used.[2,3] However, sleepiness and sleep propensity are not the only aspects of the burden of narcolepsy. An aspect that is gradually more recognised is the quality of the awake state, for which the ability to sustain attention is an important requisite. The Sustained Attention to Response Task (SART), designed to assess this function, has previously been used in narcolepsy,[4,5] and has shown clear potential to quantify the impairment in function during wake in narcolepsy.

The SART is a go/no-go task in which the no-go target appears unpredictably and rarely, and in which both accuracy and response speed, quantified as reaction time (RT), are important. The SART was developed to investigate lapses of sustained attention in individuals with neurological impairment, and proved to be a useful tool to investigate sustained attention in a number of other clinical conditions, including sleep disorders.[4-6]

To date, the validation of the SART as a tool to measure sustained attention in sleep disordered patients is based on a comparison of SART results between patients with narcolepsy and healthy controls.[4,5] The SART discriminated well between these groups, i.e. it demonstrated good construct validity. Between-subjects variability in SART performance was higher in the narcolepsy group than in the control group. No correlations were found between SART performance and subjective sleepiness (ESS) or between SART performance and the average sleep onset latency during multiple sleep latency tests (MSLT), i.e. the SART showed discriminant validity with these measures of sleepiness/sleep propensity.

As the SART quantifies the impairment of the waking condition in narcolepsy, it should also be a useful tool to measure treatment effects in narcolepsy. Hence, the objective of this study was to validate the SART as a measurement of treatment in narcolepsy, and to compare it with the MWT and ESS, two tests frequently used in treatment-effect studies in hypersomnias[7-10] that, however, have never explicitly been validated for their capability to measure treatment effects in narcolepsy. As the initial studies of the SART in sleep disorders have neither assessed the reliability of the test, nor the statistical properties of its outcome measures (i.e. descriptive statistics, statistical distribution of the data), these characteristics were also investigated in this study and compared to those of the ESS and MWT.

METHODS

Subjects

The analysis was conducted on data originating from a double-blind, parallel-group, multi-centre trial comparing the effects of eight-week treatment with the experimental drug BF2.649 (pitolisant) to effects of the proven effective drug modafinil and to placebo in narcolepsy (NCT01067222).[11] Inclusion criteria were the presence of narcolepsy with or without cataplexy diagnosed according to the International Classification of Sleep Disorders (ICSD)-2 criteria and a score of ≥14 on the Epworth Sleepiness Scale (ESS) during the baseline period.

The trial was conducted in accordance with the International Conference on Harmonization Guidelines for Good Clinical Practice and the Declaration of Helsinki. The protocol was approved by central and local ethics committees and written informed consent was obtained from all subjects prior to the study. The results of this study were published separately.

Design

Eligible patients started with a baseline period of seven days in which they were not allowed to take psychostimulants, medication with sedating properties, tricyclic antidepressants, psychoactive agents, or medication interacting with modafinil. Patients were allowed to take their anticataplectic drugs (sodium oxybate and nontricyclic antidepressants). The baseline period was completed by an inclusion visit. Patients continuing to meet the inclusion criteria were randomly assigned to one of three equally sized treatment groups for a total duration of eight weeks, with possible titration after two weeks and, if necessary, also after three weeks. A control visit took place after seven weeks, and an endpoint visit took place after the eight-week treatment period.

The SART and the MWT were performed at the inclusion visit and the endpoint visit (or the last on-study visit). A SART session was administered prior to each of four MWT sessions, starting at 10:00 hrs and at two-hour intervals thereafter. Patients were requested to take their morning treatment and to have a light breakfast before 08:00 hrs, arriving at the trial centre around 09:00 hrs. Patients took trial medication and had lunch immediately after the second MWT session. Patients were to refrain from stimulating beverages such as coffee or tea during these visits.

Sustained Attention to Response Task

The SART involved withholding key presses to 1 in 9 target stimuli during a 4-minute 19-second period. A number from 1 to 9 was shown 225 times in white on a black computer screen in a quasi-random way, while patients were seated on a chair in front of a computer screen. The font size was randomly chosen from 26, 28, 36, or 72 points. Each number was presented for 250 milliseconds, followed by a blank screen for 900 milliseconds. Subjects had to respond to the appearance of each number by pressing a button, except when the number was a 3. Subjects had to press a button before the next number appeared and were instructed to give equal importance to accuracy and speed in performing the task.[6,12]

The primary outcome measure of the SART is the total number of errors, consisting of, firstly, key presses when no key should be pressed (i.e. after a '3', a so-called 'no-go trial') and, secondly, absent presses when a key should have been pressed (i.e. after anything but a '3', the so-called 'go trials'). Errors on no-go trials, with a maximum count of 25, are called commission errors. Errors on go trials, omission errors, have a theoretical maximum count of 200. The sum of commission and omission errors, the total error count, was also analysed.

Maintenance of Wakefulness Test

The MWT consisted of four 40-minute sessions in a quiet and dimly lit room. Subjects were instructed to stay awake while comfortably seated in a semi-supine position. Movements or vocalisations were not allowed. The session was terminated either when sleep-onset occurred, defined as either three consecutive 30-second epochs of stage 1 sleep or a single 30-second epoch of any other sleep stage, or after 40 minutes of being awake.[13] The mean of the four sleep-onset latencies was considered the primary outcome measure of the MWT.

Epworth Sleepiness Scale

The ESS was administered twice at baseline (at the start of the baseline period and at the inclusion visit) and twice after treatment (at the control visit and the endpoint visit). The two early and the two late measurements were treated as separate sessions in order to assess reliability, i.e. they were not averaged. The four sessions were also separately used in the analysis of treatment efficacy.

Clinical Global Impression

The severity of EDS and of cataplexy was assessed by the local investigator using the Clinical Global Impression of Severity (CGI-S),[14] a 6-point scale, at both baseline visits. Their average value was used for analysis. Any changes in severity of EDS and of cataplexy were measured by the investigator using the Clinical Global Impression of Change (CGI-C) at each follow-up visit.[14] Ratings of this 7-point scale were averaged for the control and endpoint visit to create the final CGI-C score. CGI-S and CGI-C were rated based on a clinical interview before the administration of other scales or tests. The CGI-S and CGI-C scores were linearly transformed into a range from 0 to 4 to enhance comparability, with low values indicating higher severity in the CGI-S and more worsening in the CGI-C.

Statistical analysis

The statistical analyses were carried out with R statistical package (R, version 2.12.2). Unless specified otherwise, we conducted two-sided tests with a significance level of 0.05.

Descriptive statistics

Normality of SART, MWT and ESS outcome measures was assessed by descriptive statistics, parameters of asymmetry and kurtosis, box plots, and the Kolmogorov-Smirnov (KS) test for normality. In case of non-normality, this was repeated for the log transformation of the respective outcome measures. Floor and ceiling effects and homoscedasticity (homogeneity of variance) were tested in subgroups based on age range and gender.

Reliability

A test is considered reliable when within-patient variability is low; no significant change in the test value should occur during a period in which no change is expected, and the value should respond when such a change in condition occurs. We calculated the reliability of each outcome measure with a linear mixed model (see Appendix 6.1).

Reliability is high when within-patient variability is low compared to the variability of the studied outcome measure. To express this comparison as a number the intra-class correlation coefficient (ICC) of reliability was used.[15] The ICC was estimated from our model as follows: the within-patient variability (squared) was divided by total variability, which is the within-patient variability (squared) plus the variability of the studied outcome measure (squared).

The optimal value of the ICC is 1, meaning there is no within-patient variability, and all variability is explained by variability of the studied outcome parameter. An ICC > 0.8 is accepted as indicating good reliability.[16,17]

When one measurement or test session proves to have an insufficiently high reliability, this reliability can be increased by repeating the test.[18] As the SART and MWT were each performed four times on a test day, the ICCs resulting from the first 2 to all 4 sessions were calculated using the Spearman-Brown expression for stepped-up reliability.[19]

Sensitivity

As we aimed to investigate the validity of the SART in the context of narcolepsy, the CGI-C was considered the most appropriate standard to compare SART results with, as it reflects clinically pertinent changes in a patients' condition, assessed in a manner reflecting normal medical practice in a patient-physician interview. We calculated the sensitivity of each outcome measure for treatment efficacy by dividing subjects into responders and non-responders. Such a classification provides two groups that are supposed to differ in the true level of the constructs, but that are quite homogeneous within each group. The best dichotomy between the categories was found through assessing the linearity of the scale (Logit model between CGI-C and first factor from a confirmatory factor analysis), corroborated with a Rasch Analysis. On this basis, a responder was defined as being 'much' or 'very much' improved on the CGI-C, and all other results were classified as non-responders. This strategy is commonly used in various studies.[20-23] Analysis of covariance (ANCOVA) was used to compare outcome measures between responders and non-responders, corrected for baseline values, age, and sex.

The difference in the mean outcome measure between responders and non-responders was divided by its standard deviation (called 'residual standard deviation') to calculate the so-called Cohen's Effect Size (ES) or standardised mean difference. An ES > 0.5 is considered clinically relevant. If baseline and final values of the same outcome measures are correlated, the residual standard deviation is reduced, leading to a higher ES. A corrected effect size taking into account this correlation was calculated by multiplying the ES by the square root of the coefficient of correlation between the baseline and final values. The effect size was also measured using a linear mixed model, in which the interaction between treatment effect and time provided a more accurate measure of the effect size.

Finally, associations between CGI-C, MWT, ESS, and SART were investigated using factor analysis to demonstrate the contribution of each outcome measure to the CGI-C score.[24,25]

Missing values

Reliability and sensitivity were estimated on the available data set. The trial from which our data originated was considered a pivotal Phase III trial. As such, missing data were rare, with no missing data at baseline and less than 7% at final time. These data were not imputed, but directly handled by the mixed model.[26]

For sensitivity purposes, we repeated our analyses in imputing missing data by using Last Observed Carried out Forward techniques (LOCF), Baseline Carried Forward (BCF) and multiple imputation. We calculated the relative error between the values of the three techniques Qi with our suggested method Q in calculating E=100* | (Qi-Q)/Q |. These values were 0.8%, 1.3%, and 1.7% respectively. We therefore concluded that imputation of missing data did not change the results.

RESULTS

Subjects and data characteristics

Patient characteristics are summarised in Table 6.1. None of the SART accuracy measures was normally distributed (KS, $p<0.001$). After logarithmic transformation, the commission errors and the total number of errors became normally distributed (KS, $p=0.14$). No suitable transformation was found that resulted in a normal distribution for omission errors (KS, $p<0.001$). The ESS showed a slightly platycurtic normal distribution (KS, $p>0.55$). A ceiling effect was observed for the MWT, caused by the maximum score of 40 minutes; this made the nature of the distribution difficult to define with precision. A log-normal distribution was suspected (KS, $p=0.23$) and was therefore used in further analysis. As observed more often after log transformations, between-category heteroscedasticity was found for log-MWT.

Reliability

Table 6.2 presents within-patient variability and variability of the studied measure (i.e. the various SART error counts, ESS, and MWT) as modelled. With the aid of these estimates the ICC was calculated. The ICC was highest for the ESS at 0.83; ICC for log-SART total error

Table 6.1 Patient characteristics

	Responders (N=51)	Non-responders (N=44)	
Parameter	Mean±SD	Mean±SD	p-value
Age (years)	38.24±14.08	39.25±15.36	0.737
Sex (males (%))	28 (55%)	24 (55%)	0.971
Baseline ESS	18.70±2.79	18.13±2.39	0.291
Baseline CGI-S	1.63±1.04	0.93±0.51	<0.001
Baseline SART total errors	15.65±13.69	11.62±7.21	0.079
Baseline MWT sleep latency (min.)	10.6±8.9	13.2±10.6	0.196
Endpoint ESS	9.76±6.56	15.02±4.12	<0.001
Endpoint CGI-C	3.26±0.83	0.91±0.29	<0.001
Endpoint SART total errors	8.77±7.03	11.48±8.91	0.145
Endpoint MWT sleep latency (min.)	23.6±14.6	12.4±11.1	<0.001

ESS: Epworth Sleepiness Scale; CGI-S: Clinical Global Impression of Severity; CGI-C: Clinical Global Impression of Change; SART: Sustained Attention to Response Task; MWT: Maintenance of Wakefulness Test; log: log-transformed; min: minutes; SD: standard deviation.

Table 6.2 Variability and intra-class coefficient of correlation of SART, MWT and ESS

	Within-patient variability	Variability measure	ICC
SART commission errors (log)	0.14	0.23	0.71
SART omission errors (log)	0.30	0.34	0.56
SART total errors (log)	0.20	0.28	0.65
MWT sleep latency in min. (log)	0.26	0.47	0.76
ESS	1.09	2.45	0.85
SART commission errors	2.85	4.39	0.70
SART omission errors	8.28	7.97	0.48
SART total errors	8.67	10.50	0.59
MWT sleep latency in min.	7.44	13.48	0.77

ICC: intra-class coefficient of correlation, calculated as follows: within-patient variability (squared) divided by the total variability, which is the within-patient variability (squared) plus the variability of the studied outcome measure (squared). The last four rows illustrate that non-log-transformed SART and MWT have a lower ICC compared to their log-transformed match.

count was 0.65, and for log-MWT it was 0.76. The influence of replication is presented in Table 6.3; repeating the test improved the reliability for the MWT to 0.87 for the first two tests and to 0.82 for the first two log-transformed SART commission error counts.

Table 6.3 Influence of the number of sessions on the intra-class coefficient of correlation

	ICC	Replication		
	1	2	3	4
SART commission errors (log)	0.70	0.82	0.88	0.90
SART omission errors (log)	0.56	0.72	0.79	0.84
SART total errors (log)	0.65	0.79	0.85	0.88
MWT sleep latency in min. (log)	0.76	0.87	0.91	0.93
ESS	0.83	0.91	0.94	0.95
SART commission errors	0.70	0.83	0.88	0.90
SART omission errors	0.48	0.65	0.74	0.79
SART total errors	0.59	0.75	0.81	0.85
MWT sleep latency in min.	0.77	0.87	0.91	0.93

ICC: intra-class coefficient of correlation. An ICC > 0.80 is regarded as good reliability. The ICC resulting from the first 2 to all 4 sessions was calculated using the Spearman-Brown expression for stepped-up reliability.[19] In bold the minimum number of sessions necessary to provide good reliability (ICC>0.80).

Sensitivity

SART, ESS and MWT results differed significantly between the responder group and the non-responder group with lower SART and ESS scores and higher MWT sleep latencies for responders (Table 6.4). The corrected ES was ≥ 0.5 for all outcome measures except for the SART omission error count (Table 6.5). The highest effect size was seen for the ESS.

Using these results, we calculated which sample size would be needed to perform a treatment-effect study based on common assumptions (i.e. α=0.05, 1=β=0.9, two-sided

Table 6.4 ANCOVA non-responders versus responders corrected for age and sex

	Delta	SD	p-value
SART commission errors (log)	0.13	0.16	<0.001
SART omission errors (log)	0.17	0.31	0.007
SART total errors (log)	0.21	0.20	<0.001
MWT sleep latency in min. (log)	- 0.33	0.32	<0.001
ESS score	6.90	5.20	<0.001
SART commission errors	1.83	3.08	0.007
SART omission errors	5.30	7.77	0.002
SART total errors	7.29	8.77	<0.001
MWT sleep latency in min.	- 13.00	11.50	<0.001

Delta is calculated by subtraction of the values of the parameters of the responders from the non-responders.

test). To do so, we performed an analysis of covariance, in which we corrected for the baseline values, age and sex. As the results show (Table 6.6), using the log-transformed SART commission error count or the total error count allows studies to be designed with lower numbers of subjects than holds if non-transformed outcome parameters are used.

Table 6.5 Cohen's effect size

	Coefficient of correlation	Effect size	Corr. effect size
SART commission errors (log)	0.81	0.81	0.68
SART omission errors (log)	0.64	0.55	0.41
SART total errors (log)	0.76	1.05	0.85
MWT sleep latency in min. (log)	0.63	1.01	0.88
ESS score	0.34	1.33	1.21
SART commission errors	0.50	0.59	0.50
SART omission errors	0.50	0.68	0.47
SART total errors	0.64	0.83	0.64
MWT sleep latency in min.	0.57	1.13	0.99

Cohen's effect size (ES) was calculated by dividing the difference in the mean outcome measure between responders and non-responders by its standard deviation. The corrected effect size was calculated by multiplying the ES by the square root of the coefficient of correlation between the baseline and final values. An ES > 0.50 is regarded as good.

Table 6.6 Necessitated sample sizes

	N1	N2	N3	N4
SART commission errors (log)	16	14	13	13
SART omission errors (log)	74	58	53	50
SART total errors (log)	13	11	10	10
MWT sleep latency in minutes (log)	16	14	14	13
ESS	13	12	12	12
SART commission errors	64	54	51	50
SART omission errors	71	53	46	43
SART total errors	31	24	23	22
MWT sleep latency in minutes	15	13	13	12

log: log-transformed. N1 until N4 is the sample size necessitated in case of 1/2/3/4 tests or sessions in standard conditions (α=0.05, 1=β=0.9, two-sided test), calculated by an analysis of covariance of the values of the outcome measure, corrected for the baseline values, age and sex.

Comparison of the MWT, ESS and SART

Figure 6.1 shows the results of the factor analysis in which the CGI-C noted at the final patient visit was compared to the mean change from baseline (or relative variation) of the SART, ESS and MWT. There was a significant correlation between CGI-C and all outcome measures with the highest correlation for delta log-SART total error count (r=0.606) and the ESS (r=0.535).

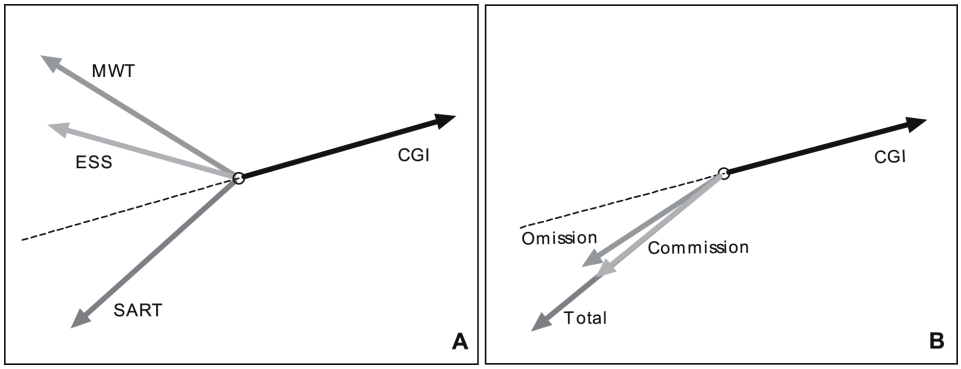


Figure 6.1 Factor analysis of CGI-C, SART, ESS, and MWT.
(A) Factor analysis of the delta scores of the MWT (log-transformed), ESS, SART (log-transformed total error count) and CGI-C. (B) Factor analysis of the delta scores of SART outcome measures (all log transformed), and CGI-C.
The direction of the arrows represents the degree of correlation between the various measures. When arrows point in the exact same direction, they are perfectly positively correlated. Arrows pointing in opposite directions, with an angle between them of 180° indicate a perfect inverse, i.e. negative correlation. Arrows at right angles to one another reflect that the two measures are completely independent. The dashed line represents the 180° opposite of the CGI score.
Figure A shows that the ESS and SART are more parallel to the CGI than the MWT, and the angle between them suggests that they capture different aspects.

DISCUSSION

This study demonstrated that the SART is a useful tool to measure treatment efficacy in narcolepsy. Of the various SART outcome measures, the log-transformed total error count proved most sensitive to treatment effects, as established by the CGI-C. The log transformed commission error count proved the most reliable across sessions performed on the same day; a good reliability of >0.8 was already reached after performing the SART twice. Performing the SART three times allowed the log-transformed total error count to exceed this threshold as well.

Reliability of the SART, ESS and MWT

Tests are considered reliable when no significant changes in their outcome measures are observed in periods when no change is expected, and when such changes do occur when there is a change in condition. We used the intra-class correlation coefficient to compare reliability of the SART, the ESS and the MWT. As the ESS needed to be administered only once to reach a high level of reliability, the ESS proved the most reliable test. Note that repeated administration of the ESS differed from repeated administration of SART and MWT: the ESS was repeated with an interval of one week, while SART and MWT sessions were repeated on the same day. However, we did not consider this a limitation of high importance, as we aimed at comparing the reliability of SART, MWT and ESS in their usual schedule of administration. The ESS measures experienced sleepiness over the past week(s) or month(s), and is therefore not administered several times per day.

The SART can achieve the same level of reliability as the ESS, but to do so it needed to be administered twice when the log-transformed commission error count was used, and three times when the log-transformed total error count was used.

The MWT reached a similar level of reliability after two sessions, regardless of log transformation. The distribution of the MWT exhibits a ceiling effect meaning that using it for statistical analysis is complex and should be treated with caution.

These results suggest that SART and MWT can measure treatment effects reliably using only the first two or three sessions on one day instead of the four sessions that are conventionally used, given the fact that they are performed at the same time of day as in this study. More than three sessions probably do not relevantly further explain variability and using four tests will be accompanied by higher costs and longer duration. Those who wish to investigate a time of day effect on treatment results might wish to use four or even more tests, but should realise that time of day (morning vs afternoon) did not affect SART performance in a recent study.[27]

Sensitivity of the SART, ESS and MWT

Note that a high reliability of a test does not necessarily mean that it also reflects clinical improvement well. We investigated the latter aspect, sensitivity, for which we used the CGI-C as a gold standard. The ESS, SART and MWT all showed high sensitivity, with highest

sensitivity for the ESS. The highest effect size of SART was found for the log-transformed total error count. We also found that the change in clinical condition from baseline to endpoint (CGI-C) was significantly correlated with the changes (delta scores) of all three tests. In fact, of the three studied measurements, the change in a SART parameter (log-transformed total errors count) reflected the change in clinical condition most closely, followed by the ESS.

Which aspect of improvement do the various tests reflect?

The SART, ESS and MWT need not reflect the same aspects of the burden of narcolepsy. In fact, in previous studies the SART error count was not related to the ESS, which reflects perceived sleepiness, and the MSLT, which reflects the propensity to fall asleep quickly.[4,5] In these same studies, ESS and MSLT results were correlated. We attempted to unravel the correlation between our outcome measures through factor analysis (Figure 6.1). The arrows of the ESS and MWT roughly point in the same direction, which means that changes in MWT and ESS during the study largely reflect the same aspect of the narcolepsy burden. Of these two measures, the treatment response as expressed in the delta ESS score is the better representative of the investigator's impression of treatment response, as the angle between CGI-C and delta ESS is smaller than between CGI-C and MWT. The delta scores of the ESS and SART explain the CGI-C score quite well (i.e. lie close to the 180° opposite of the CGI-C arrow) in a similar magnitude. Interestingly, the delta scores of the ESS and SART form a large angle (close to 90°) among themselves, indicating that they indeed explain different aspects of the CGI-C score. The factor analysis thus shows that the investigator's impression is both based on sustained attention and the ability to stay awake.

The optimal test battery to measure treatment response in narcolepsy

Measures of sleep propensity (MWT) or perceived sleepiness (ESS) on the one hand, and sustained attention on the other hand (SART), are complementary. This study indicates that a combination of the SART with either the MWT or ESS comprises the most suitable combination of the three investigated tests to measure treatment response in narcolepsy. As the MWT and the ESS in part seem to explain the same variability, the question rises which of these tests is best suited to measure treatment effects. The MWT and the SART measure more distinct phenomena than the ESS and SART, as the angle between delta-MWT and delta-SART is closer to 90° in the factor analysis. Another argument in favour of the MWT is that it is easier for a patient to manipulate an ESS result for whatever reason than an MWT

result. Then again, there are a number of arguments against the MWT. It can be manipulated by reducing previous amount of sleep; it is not uniformly carried out: some use 20-minute sessions, others 40-minute ones; some use four, others five sessions; the definition of sleep onset also varies. However, these disadvantages could be overcome by using the protocol recommended in the AASM manual.[13] Furthermore, the MWT is performed in an artificial setting that need not represent daily life, and, finally, it is time-consuming. Compared to the MWT the ESS is inexpensive, has a high degree of internal consistency and can easily be re-rated for follow-up studies. While these arguments were already known the present study adds new ones in favour of the ESS over the MWT: it had the highest reliability of all three tests and was more sensitive to treatment efficacy than the MWT.

An interesting characteristic of the SART has to do with the balance between a subjective and objective assessment. An 'objective' test reflects a quantitative test measurement rather than a patient's opinion of disease severity. The SART (and MWT) as objective tests offer the advantage of immunity to manipulation in one direction: it is possible to perform the test worse than one's conditions allows, but not better. However, 'subjective' assessment by patients often forms the primary reason to alter treatment in patient care. The SART has the advantage of objectivity as well as a close relation to subjective changes in severity, reflected in the CGI-C.

We conclude that a single ESS accompanied by two to three SART sessions, depending on the chosen SART outcome parameter, provides a good method to evaluate treatment effects in narcolepsy. This battery comprises two key aspects of narcolepsy, perceived sleepiness and sustained attention, and is easy and cheap to administer.

Detailing SART analysis

Which SART parameter should be used? The factor analysis revealed only minor differences among the various outcome measures of the SART (Figure 6.1b), indicating that they represent the same part of the CGI-C. The highest effect size was found for the total error count. This needs three SART sessions, compared to two for the commission error count. The latter parameter also had a better distribution. The omission error count did not perform as well in terms of distribution and reliability. Still, the total error count did perform well, and, as it contains the omission error count as well, counting omission errors may have a role. The relative importance of omission, commission and total error counts can differ

between disorders.[28] We accordingly advise to use the total error count as the primary SART outcome measure.

Reaction time can also be used as a SART parameter, but measuring RT accurately requires special equipment, whereas measuring error counts can be done with standard personal computers. In the present multicentre study, RTs were not measured.

A different test to measure sustained attention is the Psychomotor Vigilance Task (PVT), which has been used and validated in sleep deprivation studies.[29-31] PVT results in narcoleptics differed from those of healthy controls.[32] The PVT is sensitive to treatment efficacy in obstructive sleep apnoea syndrome,[33] but its role in assessing treatment efficacy in narcolepsy would require an assessment similar to the present study, which is currently not available.

Study limitations

Our study is limited to the three measurements of treatment effect that we evaluated, so there is no way for us to tell whether any of the other possible parameters to measure disturbed sleep and its consequences would be useful to measure treatment efficacy.

Our results are based on data from a study designed to evaluate effects of pitolisant. This means that patients were not selected to represent a typical spectrum of severity of narcolepsy. However, the selection was not limitative, the statistical analysis was prepared before the analysis of the drug trial, and the analysis was conducted independent of the main and secondary endpoints of the trial.

Conclusion

In conclusion, this study shows that the SART, in particular the commission errors and the total error score, is a valid measure to detect treatment effects. A combination of the SART and ESS includes a comprehensive evaluation of treatment effects in narcolepsy since the ESS represents a subjective estimate of how sleepy patients feel, while the SART is objective in nature. Together they share the advantages of not requiring much time or money, and they correlate well with the clinical global assessment of patient improvement.

REFERENCES

1. Dodel R, Peter H, Spottke A, et al. Health-related quality of life in patients with narcolepsy. Sleep Med 2007;8:733–741.

2. Mitler MM, Gujavarty KS, Browman CP. Maintenance of wakefulness test: a polysomnographic technique for evaluation treatment efficacy in patients with excessive somnolence. Electroencephalogr Clin Neurophysiol 1982;53:658–661.

3. Johns MW. A new method for measuring daytime sleepiness: the Epworth sleepiness scale. Sleep 1991;14:540–545.

4. Fronczek R, Middelkoop HA, van Dijk JG, et al. Focusing on vigilance instead of sleepiness in the assessment of narcolepsy: high sensitivity of the Sustained Attention to Response Task (SART). Sleep 2006;29:187–191.

5. Van Schie MKM, Thijs RD, Fronczek R, et al. Sustained attention to response task (SART) shows impaired vigilance in a spectrum of disorders of excessive daytime sleepiness. J Sleep Res 2012;21:390–395.

6. Robertson IH, Manly T, Andrade J, et al. "Oops!": performance correlates of everyday attentional failures in traumatic brain injured and normal subjects. Neuropsychologia 1997;35:747–758.

7. Arand D, Bonnet M, Hurwitz T, et al. The clinical use of the MSLT and MWT. Sleep 2005;28:123–144.

8. Black J, Houghton WC. Sodium oxybate improves excessive daytime sleepiness in narcolepsy. Sleep 2006;29:939–946.

9. Broughton RJ, Fleming JA, George CF, et al. Randomized, double-blind, placebo-controlled crossover trial of modafinil in the treatment of excessive daytime sleepiness in narcolepsy. Neurology 1997;49:444–451.

10. Johns MW. Reliability and factor analysis of the Epworth Sleepiness Scale. Sleep 1992;15:376–381.

11. Dauvilliers Y, Bassetti C, Lammers G-J, et al. Pitolisant versus placebo or modafinil in patients with narcolepsy: a double-blind, randomised trial. Lancet Neurol 2013;12:1068–1075.

12. Manly T, Robertson IH, Galloway M, et al. The absent mind: further investigations of sustained attention to response. Neuropsychologia 1999;37:661–670.

13. Littner MR, Kushida C, Wise M, et al. Practice parameters for clinical use of the multiple sleep latency test and the maintenance of wakefulness test. Sleep 2005;28:113–121.

14. Guy W. ECDEU Assessment Manual for Psychopharmacology. Rockville, MD.: U. S. Dept. of Health, Education, and Welfare, Public Health Service, Alcohol, Drug Abuse, and Mental Health Administration, National Institute of Mental Health, Psychopharmacology Research Branch, Division of Extramural Research Programs; 1976.

15. Ferketich S. Focus on psychometrics. Internal consistency estimates of reliability. Res Nurs Health 1990;13:437–440.

16. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika 1951;16:297–334.

17. Lohr KN, Aaronson NK, Alonso J, et al. Evaluating quality-of-life and health status instruments: development of scientific review criteria. Clin Ther 1996;18:979–992.

18. Fleiss JL. The Design and Analysis of Clinical Experiments. New York: John Wiley and Sons; 1986.

19. Lord FM, Novick MR. Statistical theories of mental test scores. Reading, MA: Addison-Welsley Publishing Company; 1968.

20. Bastyr EJI, Price KL, Bril V. Development and validity testing of the neuropathy total symptom score-6: questionnaire for the study of sensory symptoms of diabetic peripheral neuropathy. Clin Ther 2005;27:1278–1294.

21. Garralda ME, Yates P, Higginson I. Child and adolescent mental health service use. HoNOSCA as an outcome measure. Br J Psychiatry 2000;177:52–58.

22. Virues-Ortega J, Carod-Artal FJ, Serrano-Duenas M, et al. Cross-cultural validation of the Scales for Outcomes in Parkinson's Disease-Psychosocial questionnaire (SCOPA-PS) in four Latin American countries. ValueHealth 2009;12:385–391.

23. Nelson E, Wasson J, Kirk J, et al. Assessment of function in routine clinical practice: description of the COOP Chart method and preliminary findings. J Chronic Dis 1987;40 Suppl 1:55S–69S.

24. Pinneau S, Newhouse A. Measure of invariance and comparability of factor analysis for fixed variables. Psychometrika 1964;29:271–281.

25. Lee H, Comrey A. Distortions in a commonly used factor analytic procedure. Multiv Behaviour Research 1979;14:301–321.

26. Chakraborty H, Gu H. A Mixed Model Approach for Intent-to-Treat Analysis in Longitudinal Clinical Trials with Missing Values. RTI Press; 2009.

27. Van Schie MKM, Alblas EE, Thijs RD, et al. The influences of task repetition, napping, time of day, and instruction on the Sustained Attention to Response Task. Journal of Clinical and Experimental Neuropsychology 2014;36:1055–1065.

28. Hart EP, Dumas EM, Reijntjes RH, et al. Deficient sustained attention to response task and P300 characteristics in early Huntington's disease. J Neurol 2012;259:1191–1198.

29. Dinges DF, Powell JW. Microcomputer analyses of performance on a portable, simple visual reaction task during sustained operations. Behav Res Meth Instr Comp 1985:652–655.

30. Loh S, Lamond N, Dorrian J, et al. The validity of psychomotor vigilance tasks of less than 10-minute duration. Behav Res Methods Instrum Comput 2004;36:339–346.

31. Basner M, Dinges DF. Maximizing sensitivity of the psychomotor vigilance test (PVT) to sleep loss. Sleep 2011;34:581–591.

32. Prasad B, Choi YK, Weaver TE, et al. Pupillometric assessment of sleepiness in narcolepsy. Front Psychiatry 2011;2:35–35.

33. Dinges DF, Weaver TE. Effects of modafinil on sustained attention performance and quality of life in OSA patients with residual sleepiness while being treated with nCPAP. Sleep Med 2003;4:393–402.

APPENDIX 6.1

Reliability

We calculated the reliability of each outcome measure as the ratio of its observed variability divided by the variability of the true value of the construct that was measured. As this true value cannot be measured directly, it was estimated from our data by means of a linear mixed model. We defined the following linear mixed model to compare the reliability of SART accuracy measures, MWT sleep latency and ESS score:

$$Y(i) = K + Time * [1+N(0, \sigma e)] + age + sex + \sigma T$$

In this model, we assumed that the value of the outcome measure (Y) depended on a constant value (K), the variability of the studied outcome measure (σT), some random variability expressed as the interaction of within-patient variability (σe) with time, and the effects of age and sex. The model contained a random factor for the short time interval in which the value of the outcome measure was not expected to vary within each subject.

Sensitivity

Effect size was also measured using a linear mixed model assuming that the value of the outcome measure depended on a constant value, time, being a responder or not, the interaction of the latter two, age, and sex:

$$Y() = K + Time + Responder + Time * Responder + age + sex$$