Sleep, 17(2):160–167

Clinical Research

The Sleep Disorders Questionnaire I: Creation and Multivariate Structure of SDQ

*Alan B. Douglass, †Robert Bornstein, ‡German Nino-Murcia, §Sharon Keenan, ¶Laughton Miles, **Vincent P. Zarcone, Jr., §§Christian Guilleminault and §§William C. Dement

**Ann Arbor VA Medical Center Psychiatry Service and University of Michigan Department of Psychiatry, Ann Arbor, Michigan, U.S.A.;*
†Department of Psychiatry, Ohio State University, Columbus, Ohio, U.S.A.;
‡Director, Sleep Medicine and Neuroscience Institute, Palo Alto, California, U.S.A.;
§Director, School of Sleep Medicine, Palo Alto, California, U.S.A.;
¶Ambulatory Monitoring Institute, Palo Alto, California, U.S.A.;
***Department of Psychiatry, Stanford University Medical School and Palo Alto VA Medical Center, Palo Alto, California, U.S.A.; and*
§§Sleep Disorders Center, Stanford University Medical School, Palo Alto, California, U.S.A.

Summary: The development of the Sleep Disorders Questionnaire (SDQ) from the Sleep Questionnaire and Assessment of Wakefulness (SQAW) of Stanford University is described in detail. The extraction of the best question items from the SQAW and their subsequent rewording in the SDQ to insure greater completion rates are described. Two item test–retest reliability studies are reported on 71 controls and on 130 sleep-disorder patients, which confirmed adequate reliability. To create multivariate scoring scales, SDQ was then given in a multicenter study to 519 persons, 435 of whom were sleep-disorder patients with full polysomnography. Canonical Discriminant Function Analysis was employed, which resulted in four clinical-diagnostic scales: SA for sleep apnea, NAR for narcolepsy, PSY for psychiatric sleep disorder and PLM for periodic limb movement disorder. Each was adjusted for male and female responses and transformed to a percentile using the observed distribution of raw scores. Using Receiver Operating Characteristics analysis, cutoff points were determined for each scale to maximize its sensitivity and specificity. Positive and negative predictive values were also calculated. The SA and NAR scales proved to be the most discriminating. **Key Words:** Questionnaire—Multivariate scales—Narcolepsy—Sleep apnea—Sensitivity—Specificity—Receiver operating characteristics.

The field of sleep-disorders medicine is relatively new. Its progress can be attributed to the objective diagnosis of sleep pathophysiologies made possible by the nocturnal polysomnogram (NPSG) and the multiple sleep latency test (MSLT). Patient questionnaires about sleep habits and symptoms have been less focused upon, although many sleep laboratories have developed one that they use for various purposes, such as a data base of symptoms, a research tool, clinical documentation, a teaching aid or as a screening tool for referrals to the laboratory.

Despite the wide availability of sleep disorder centers in the U.S.A., the diagnostic procedure is expensive ($1,000 up). In many other countries sleep laboratories are either not available or are not reimbursed by government health plans. General practitioners are increasingly aware of sleep disorders, but they often identify more suspected patients than they can reason-

Accepted for publication October 1993.
Address correspondence and reprint requests to Alan B. Douglass, M. D., Psychiatry Department, Room 2951 CFOB, University of Michigan, Ann Arbor, MI 48109-0704, U.S.A.

ably refer for a full sleep-laboratory study. There is, therefore, a need for a "triage questionnaire" able to distinguish a patient at high risk from a larger group that the general practitioner believes possesses some of the symptoms of sleep disorder. Few available sleep questionnaire instruments have this diagnostic point of view. One exception is an inventory designed to assess the presence of sleep apnea by self-report (1).

Many existing instruments, such as the St. Mary's Sleep Questionnaire (2), address subjective dimensions of the previous night of sleep, such as "sleep quality" and "sleep satisfaction". Some, like Buysse's Pittsburgh Sleep Quality Index (3) elicit subjective reports about sleep that have been validated by polysomnography. The Buysse paper also contains a thorough review of other sleep questionnaires that will not be mentioned further here. There are instruments that deal with psychological symptoms surrounding sleep (4,5). Others are special-use questionnaires for daily symptom ratings or research, such as the Stanford Sleepiness Scale (6), the Leeds Sleep Evaluation Questionnaire (7) and a "morningness-eveningness" instrument (8) for use in circadian rhythm studies. A related instrument is the sleep diary, which Carskadon has found especially useful for chronobiological studies (9).

A general clinical questionnaire with wide usage is the Sleep Questionnaire and Assessment of Wakefulness or SQAW (10), written by several of the present authors and used since the 1970s at Stanford University. The present report describes the derivation of a new questionnaire (11), the Sleep Disorders Questionnaire (SDQ), from the existing SQAW.

The goals of the SDQ project are multiple. We wish to (a) recast the clinical experience gained with the SQAW using modern multivariate methods; (b) create a uniform database of clinical responses that could be used by a sleep clinic for chart documentation, augmented history-taking, outcome research or differential diagnosis; (c) estimate the chance of a patient on a waiting list actually having a sleep disorder diagnosable by polysomnogram and (d) create a pool of items from which a small subset could be used as a screening test for sleep disorders in the general public. In this report, only goals (a), (b) and (c) are discussed. Goal (c) has not yet been confirmed by replication. Goal (d) is a future development of SDQ that may depend on the success of the earlier goals.

An item test–retest reliability (consistency) study of SDQ in controls has already been published as an abstract (12). In summary, 71 persons without sleep complaint were administered the SDQ twice over a 2-week interval. Mean age of the subjects was 24.8 years, SD = 8.3. Of these subjects, 57 were college undergraduate students and 14 were psychotherapy outpatients. The item reliabilities (Pearson correlation) ranged from $r = 0.999$ to 0.163; mean $r^2 = 0.495$ (representing a mean $r = 0.704$). All except three items achieved correlations that were significant at $p < 0.0001$. The completion rate was 95.7% of items.

The approach used in the present multicenter project was to reword the SQAW's questions and derive diagnostic scales. Only patients who were diagnosed as having sleep disorders on NPSG and MSLT formed the groups upon which the scales were based. Included below are descriptions of item selection, item test–retest reliability, scoring scale derivation by multivariate methods and reliability of the derived multivariate scales. Finally, the results of a Receiver Operating Characteristics (ROC) analysis of the sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) of each scale are presented.

METHODS

Item selection for the SDQ

Several properties of the SQAW items prevented a multivariate analysis, and the SQAW also had a poor completion rate among clinical patients. We selected items from the SQAW for retention in the SDQ by applying four selection criteria: (a) high completion rate; (b) significant difference on univariate ANOVA or chi-square analysis between the clinical groups narcolepsy, sleep apnea, periodic limb movement disorder (PLMD) and psychiatric sleep disorder; (c) adequate face validity upon review by three accredited polysomnographers and (d) describes a pathognomonic symptom of any major sleep disorder, even if of low frequency in the population. In addition, factor analyses on six homogeneous subgroups of SQAW items were performed, and items with the highest loading on the two factors with highest eigenvalues were also included in the SDQ if they had not already met the above criteria. Six new items not present in the SQAW, regarding sleep apnea and sleep interruption, were written.

The SQAW items were then totally reworded into approximately eighth grade reading-level before being included in the SDQ. We chose a consistent five-level response format for all the items and arranged the wording so that higher numerical scores always reflected greater clinical severity of symptom, which was not the case on the SQAW. The 175-item SDQ was completed in 1986. Items retained from SQAW as (reworded) SDQ items have been published (11). Body Mass Index (item 176) is a special case. It is not present as a numbered item on the SDQ, but rather is calculated from patient data: (weight in kg)/(height in meters)^2.

Item test-retest reliability study of SDQ in sleep-disorder patients

This preliminary study is reported here for the first time. Its purpose was to insure that the item reliabilities observed in normals were replicable in actual sleep-disorder patients. The patients were expected to have lower average educational levels than the college students, to be older and perhaps have impaired reading concentration due to their illness.

One hundred thirty patients who were referred to a sleep disorders center by their general practitioner were given the SDQ twice while on a waiting list: once by mail upon initial telephone contact and again 3–4 months later when they were studied in the sleep laboratory. There were 85 males and 45 females, mean age 43.6 years (SD = 12.9). Both administrations of the SDQ occurred before diagnosis by polysomnography. There was no control or monitoring of the treatment the referring physician provided in the interim, which may have caused a change in SDQ responses. No sampling technique was used. These were the first 130 sequential patients from the waiting list who were found to have a sleep disorder by polysomnography.

To measure item reliability over the time interval, Spearman rank correlation *rho* was used for the five-level items (1–152), and Pearson correlation was used for numeric items 153–175. Completion rate was 93.8% of items. Range of correlation was *rho* or $r = 0.308$–0.985 (all significant at $p < 0.0001$); mean $r^2 = 0.404$ or mean $r = 0.636$. There were 28 items that achieved an r or *rho* ≥ 0.80. The pattern of highly reliable items resembled that of the normal controls, except for some symptoms pathognomonic of the major sleep disorders. In the latter, the patient group showed a higher reliability than the controls.

Multivariate analysis of SDQ

After the patient had completed the SDQ and polysomnography, the clinical chart primary diagnosis was used to place him in a diagnostic group. Both primary and secondary diagnoses were allowed. Multivariate statistics based upon primary group classification were then used to select those SDQ items that most strongly predicted group membership, thus creating scoring scales.

Subjects and methods

The study involved the responses of 519 persons, 435 of whom were clinical sleep-disorder patients. There were five groups: sleep apnea (APNEA), narcolepsy (NARCO), inpatient psychiatric (PSYCH), nocturnal myoclonus/PLMD (MYOCL) and normal controls (CONTR). See Table 1 for the size of each group and ages.

TABLE 1. *Experimental groups subdivided by sex*

Group	Males			Females			Total (n = 519)
	n	Mean age	SD	n	Mean age	SD	
Sleep apnea	141	49.2	8.4	17	50.9	11.1	158
Narcolepsy	39	34.9	5.6	34	37.1	11.4	73
Psychiatric	47	32.3	16.1	61	32.6	13.9	108
PLMD	66	47.3	9.8	30	46.7	9.4	96
Normal control	39	28.4	11.3	45	25.4	9.1	84

All subjects were diagnosed using polysomnography, except 71 of the 84 controls. The control group consisted of 57 of the 71 controls mentioned above. (The 14 psychotherapy patients mentioned in test–retest study number 1 were removed from the control group and replaced by 14 hospital workers to avoid confounding psychiatric illness in the control group.) An additional 13 hospital workers were recruited by advertisement as additional normal controls. All of the latter had full NPSG studies that showed normal sleep. Final sleep-disorder diagnoses were made by the patient's treating sleep clinician, based upon the case history and polysomnography, according to the Diagnostic Classification of Sleep and Arousal Disorders (13). In the case of the psychiatric patients only, all were additionally diagnosed using the Schedule for Affective Disorders and Schizophrenia (14). Thirty of the psychiatric patients had schizophrenia; the remaining 78 had a major depression. It was not possible to screen the controls for psychiatric illness or with a physical examination. The 130 sleep-disorder patients from the test–retest study were included among the 435 sleep-disorder patients.

To reduce variance, only one hospital in the multicenter project provided the members for a given diagnostic group. This meant that all members of a given group were seen in diagnostic interview by the same group of sleep clinicians, studied in the same sleep laboratory and scored by the same technicians. To avoid bias in the clinician's diagnosis all entries on the clinical chart including primary and secondary diagnoses were completed before the multivariate scale scores were calculated. The following numbers of patients had only a single diagnosis: apnea (B4), 129; narcolepsy (B6), 30; psychiatric (A2), 103 and PLMD (B5), 68. Twenty-seven apnea patients had a secondary diagnosis of PLMD. Secondary diagnoses among narcolepsy patients included apnea, 14, and PLMD, 24. Five psychiatric patients had apnea. Secondary diagnoses among PLMD patients included apnea, 21; central nervous system hypersomnolence (B7), 2 and psychiatric, 4. The only tertiary diagnoses recorded were five narcoleptics with PLMD.

TABLE 2. *Results of canonical discriminant function analysis (CANDISC)*[a]

Variable	Canonical *r*	SE	r^2	Eigenvalue	Cumulative variance explained	Likelihood ratio (*F*)	df	p
1	0.847	0.012	0.716	2.53	0.62	6.00	256	<0.0001
2	0.683	0.023	0.466	0.87	0.83	3.53	189	<0.0001
3	0.538	0.031	0.289	0.41	0.94	2.42	124	<0.0001

[a] The first three significant canonical variables from the CANDISC are shown. Raw data consisted of 64 items from SDQ measured on 519 subjects. Significance of the CANDISC multivariate procedure: Wilk's Lambda = 0.085, $F = 6.01$, df = 256, $p < 0.0001$; Pillai's Trace = 1.68, $F = 5.12$, df = 256, $p < 0.0001$.

Polysomnography

The standard NPSG monitoring included electroencephalogram (C3, C4, O1, O2, A1, A2 leads), electrooculogram by disk electrodes at the outer canthi, nasal airflow by thermistor, electromyogram of chin and anterior tibialis muscles, inductance plethysmography of the chest and abdomen, oxygen saturation by finger oximetry, and electrocardiogram "lead II". If there was a suspicion of narcolepsy from the clinical interview, an MSLT was performed the next day using a standard protocol (15). The NPSG and MSLT are not reported in detail here. They can be found in a comparison article ("SDQ II").

Statistical methods

(1) multivariate statistical design. To derive scoring scales, 64 SDQ items with test–retest reliability $r \geq 0.70$ were chosen for entry into a Canonical Discriminant Function analysis (CANDISC program, SAS version 6.03, SAS Institute, Cary, NC). Due to the different pattern of item reliability in the normal controls versus the sleep-disorder patients, an item was selected if it achieved $r \geq 0.70$ in either subject pool.

CANDISC was selected because of the nature of the multivariate sample. Rather than a homogeneous subject sample, which could be assumed to have a multivariate normal distribution, the present sample was assembled from five separate populations—four patient groups and normals. Accordingly, a single factor analysis was inappropriate. Discriminant function analysis would have provided a weighted sum of item scores for each group, but we wished to derive normally distributed clinical scoring scales for each diagnosis. The CANDISC procedure created canonical variables from the items of the SDQ. Three significant canonical variables were found. The "between-groups" loading of each item on the canonical variables was calculated. Each had zero as a center point, with positive and negative "between groups" item-loadings on either side. The SDQ items were found to cluster away from the zero point on all except the third canonical variable. The tips of canonical variable vectors 1 and 2 were homogeneous and were interpretable as three of the specific diagnostic scales: sleep apnea (SA), narcolepsy (NAR) and psychiatric sleep disorder (PSY). Items for the last scale (PLM) came from canonical variable 3.

To qualify tentatively for scale membership, an item had to have a "between groups" loading of ≥ 0.80. In the case of an item having similar loadings on more than one diagnostic scale, it was included only in the scale upon which it achieved the highest mean score.

(2) Internal consistency of scales. The four preliminary scales were submitted to a Cronbach's Alpha procedure, n = 519, to test for scale homogeneity. Several items were deleted at this point due to lack of consistency with their scale. Each item on each scale was then correlated with its scale total, minus that item.

(3) Scoring calculations. Due to the high completion rate of the SDQ, a procedure for filling in the small number of missing data was created as follows: a stratum mean value was calculated for males and females for each of the five groups in the analysis, giving 10 possible values for each of the 175 items. Missing data were then replaced by the appropriate mean, stratumwise.

Once the item membership for the four scales was finalized, scale scores were calculated for each subject by summing the "1 . . . 5" item scores of the scale's items. No weighting factor was used.

In all versions of SDQ before version 2.00, the numeric items 153–175 were written by the subject in a blank space with suggested units of measurement (e.g. "__ inches" for height). This manner of response is time-consuming to encode for computer analysis, so the 519 subjects were used to estimate an empirical distribution of the numerical scores. An SAS routine ("proc UNIVARIATE") created "quintiles" of the observed cumulative score distribution. When this approach was incorporated into SDQ (version 2.00, 1991) the response set for the whole questionnaire became a "1 . . . 5" numeric choice, facilitating subject compliance.

(4) Scale test–retest reliability and validity. To assess the test–retest reliability of the four new scales, data from the item test–retest reliability study with 130 sleep-

TABLE 3. *Scale membership of items and correlation to total scale*

Diagnostic scale	Item	*r*	Description
SA	Q21	0.71	Snore that bothers others
	Q22	0.70	Stop breathing in sleep
	Q23	0.51	Awake unable to breathe
	Q25	0.19	Sweat at night
	Q71	0.38	High blood pressure (history)
	Q139	0.38	Nose blocks up while trying to sleep
	Q141	0.67	Snoring/breathing worse if on back
	Q142	0.52	Snoring/breathing worse with alcohol
	Q163	0.62	Current weight
	Q170	0.43	Number of years as a smoker
	Q173	0.55	Age
	Q176	0.65	Body Mass Index
			alpha = 0.855
PLM	Q4	0.49	Wake up often during night
	Q12	0.50	Restless legs as falling asleep
	Q24	0.31	Palpitations at night
	Q31	0.53	Restless legs disturb sleep
	Q45	0.35	Insomnia
	Q80	0.28	Lessening of sexual desire/interest
	Q108	0.20	Smoking two hours before bedtime
	Q154	0.32	Length of longest wake period at night
	Q155	0.32	Night urination (number of times)
			alpha = 0.695
PSY	Q3	0.54	Trouble getting to sleep
	Q6	0.55	Racing thoughts at bedtime
	Q7	0.69	Sad/depressed at bedtime
	Q33	0.66	Sadness/depression disturbs sleep
	Q38	0.45	A lot of nightmares
	Q43	0.41	Unable to sleep for days
	Q84	0.48	Unhappy with loving relationships
	Q89	0.41	Considered/attempted suicide
	Q101	0.23	Family: psychiatric hospitalization
			alpha = 0.800
NAR	Q11	0.41	Feel paralyzed as falling asleep
	Q39	0.51	Paralyzed after a nap
	Q40	0.46	Hallucinations upon awakening
	Q42	0.46	Slept for several days
	Q55	0.60	Sleepy during the day
	Q56	0.47	Accidental sleep
	Q57	0.44	Bad grades due to sleepiness
	Q58	0.57	Trouble on job due to sleepiness
	Q59	0.51	Too sleepy to drive
	Q60	0.50	Hallucinations after napping
	Q62	0.53	Paralyzed upon morning awakening
	Q63	0.46	Failure to remember driving
	Q66	0.50	Weak knees when laughing
	Q67	0.49	Muscular weakness if strong emotion
	Q156	0.34	Work accidents due to sleepiness
			alpha = 0.853

A brief description of each SDQ question item is shown, not the actual item wording. "Alpha" is Cronbach's Alpha, a measure of scale consistency. SA = sleep apnea scale, NAR = narcolepsy scale, PSY = psychiatric sleep disorder scale and PLM = periodic limb movement disorder scale. All correlations are Pearson correlations and were done with the item temporarily removed from scale.

disorder patients were used. Subjects this time were scored using the four scales, and the observed scale scores over a 4-month interval were then correlated using Spearman's *rho*.

Diagnostic scales should be independent of one another unless the diseases they purport to diagnose overlap. This was tested by intercorrelating the four scales.

The observed frequency distributions of scale scores were tested against an assumption of Gaussian normality using Wilk's "W test" (16).

The means of the four scales in the five experimental groups were calculated. As a further test of internal validity, a multivariate analysis of variance (MANOVA) followed by univariate ANOVA was performed on the four scales to confirm that the five groups demonstrated significant differences from one another. This should always be the case if valid scales were created by the CANDISC/Cronbach's Alpha/item-deletion procedure.

(5) Sensitivity and specificity. A preliminary assessment of these psychometric properties was obtained using the original 519 subjects, although confirmation with a new sample needs to be done. Sensitivity and specificity for each of the four SDQ scales were simultaneously maximized, using the ROC technique (17) to set the "cutpoint" (scale score that marks the border between normal and abnormal). Sensitivity is the extent to which a scale detects patients with the target illness, whereas specificity is the extent to which it identifies normals as not having the illness. PPV and NPV were also calculated and may have more clinical utility because they bring the observed prevalence of the illness into the calculation.

(6) Criterion validity. Finally, a test of external or criterion validity of the scales was required. This is best accomplished by relating the NPSG and MSLT results of each group to their scores on the four SDQ scales (e.g. apnea index for apnea patients, MSLT sleep latency for narcoleptics, etc.). Higher scale scores should predict higher amounts of pathology on polysomnography. These results will be presented in two future articles (SDQ II, SDQ III).

RESULTS

The CANDISC procedure produced four significant variables; the first three were used to create scales. Table 2 shows these results. Although significant, variable 3 had a low eigenvalue. Its loading related to identification of PLMD symptoms, which proved to be the least powerful scale.

The test–retest reliability of the four scales in 130 sleep-disorder patients over 4 months was as follows (Spearman *rho*): SA 0.842, NAR 0.753, PSY 0.848 and PLM 0.817. All were significant at $p < 0.0001$.

Final scale membership of SDQ items is shown in

Table 3. Each item ("Q") is shown beside its correlation to the total scale value minus that item. It has been suggested (18) that such correlations should fall between 0.2 and 0.4 so that all items are not asking the same question in a different way but, rather, tapping different aspects of the condition. Cronbach's Alpha is shown at the bottom of each scale, indicating greater scale homogeneity as it approaches 1.00.

The intercorrelations of the four scales were SA–NAR = 0.14, SA–PSY = −0.20, SA–PLM = 0.34, NAR–PSY = 0.27, NAR–PLM = 0.38 and PSY–PLM = 0.48. All were significant at $p < 0.001$. Because variance explained is proportional to r^2, the intercorrelations show little connection between the scales except for the PSY and PLM scales, which have a commonality of about 25% of the variance. Also, the PLM scale persistently shows the highest correlation with other scales, suggesting that it might reflect a general condition such as "results of sleep interruption" rather than being specific to PLMD.

The empirical distributions of scale scores are shown graphically in Fig. 1. The first graph shows only members of a single patient group, subdivided by sex, on the SDQ scale diagnostic of their condition. In the second part of Fig. 1, the results from the CONTR group for all four SDQ scales are plotted in a similar manner. The greatest separation of the patient groups from controls is seen in the SA and NAR scales.

The scales achieved relatively normal distributions when scored separately by group and by sex. Wilk's W (a test of normality that approaches W = 1.00 in a true Gaussian normal distribution) varied from a low of 0.840 among male controls on the NAR scale to a high of 0.971 among male sleep apnea patients on the SA scale. When patient groups were scored on the scale appropriate to their diagnosis, the lowest was W = 0.92 in female PLMD patients on the PLM scale.

The ROC analysis results are shown in Table 4. Sensitivity and specificity of the SA, NAR and PSY scales were higher than the PLM scale. NPV was uniformly high on all scales. As expected, PPV varied strongly with prevalence, despite similar sensitivity and specificity. This is best illustrated by the male versus female apnea patients.

The post hoc MANOVA to assess internal validity was significant (Table 5). Four subsequent one-way ANOVAs are shown, one for each scale across clinical groups. Bonferroni post hoc pairwise comparisons of means demonstrated that the "characteristic" scale for a group had a significantly higher mean in that group compared to all other groups in every case except the PLM scale. In the latter, the only significant difference was between the CONTR and MYOCL groups. The PLM scale mean in the CONTR group, however, differed significantly from all four patient groups.

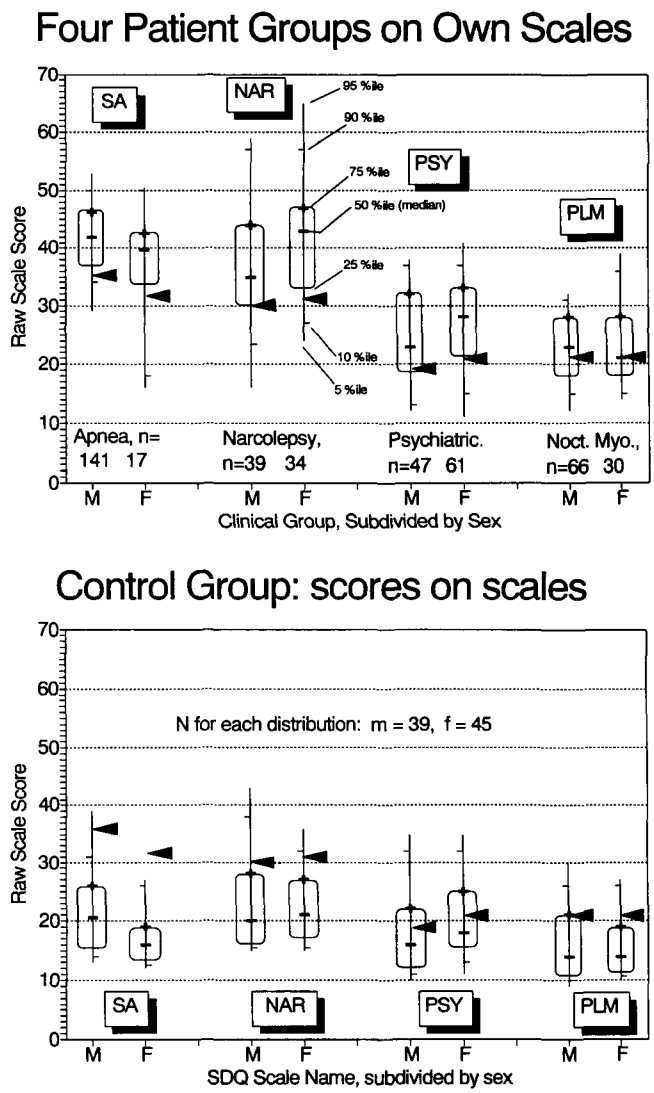

FIG. 1. Distribution graphs of SDQ scales. Observed distributions of SDQ scale scores, total n = 519. *Upper panel:* members of single patient groups, by sex, on the SDQ scale diagnostic of their condition. Percentile (%ile) ≤5 means that a patient had few or milder symptoms in common with other patients who received the same polysomnographic diagnosis. Percentile ≥95 means that a patient had all the symptoms of the illness shared by other patients with this diagnosis, and symptoms were in the greatest severity. *Lower panel:* CONTR group only. Scores shown for all four SDQ scales. Percentiles are coded in the same way on the box-whisker plots of both panels. Percentiles in the lower panel refer to comparisons within the CONTR group only.

DISCUSSION

The goal of creating diagnostic SDQ scoring scales for several major sleep disorders receives preliminary support from these results. SDQ scale scores best distinguished the APNEA and NARCO groups from the other groups in this sample. The PLM scale showed the poorest sensitivity, specificity, PPV and NPV. The PPVs and NPVs in this study are lower than they might

TABLE 4. *Sensitivity and specificity analyses*

	SDQ scale name[a]							
	SA		NAR		PSY		PLM	
	M	F	M	F	M	F	M	F
Sample prevalence (%) of scale's target diagnosis[b]	42	9	12	18	14	32	20	16
Scale cut-off point (by ROC)	36	32	30	31	19	21	21	21
Sensitivity	0.85	0.88	0.84	0.80	0.79	0.79	0.67	0.65
Specificity	0.76	0.81	0.68	0.72	0.65	0.64	0.46	0.49
Positive predictive value	0.72	0.31	0.26	0.38	0.28	0.51	0.23	0.19
Negative predictive value	0.87	0.99	0.97	0.94	0.95	0.87	0.86	0.99

These data come from an ROC analysis of each scale, by sex, on the full 519 subjects including controls. The cutoff points in row 2 are also shown in Fig. 1 as black arrow heads.

[a] M = male; F = female.

[b] Prevalence percentages should add to 100% for males, 100% for females. They do not because the normal controls are not shown as a "target diagnosis".

have been if pure diagnostic groups had been available. For example, some narcolepsy and PLMD patients had apnea. Because this is the first clinical test of SDQ, the suggested scale cutpoints, PPVs and NPVs should be regarded as provisional, pending confirmation in larger-scale studies with more sophisticated methodology, as described below.

There were limitations of the present study due to the number of subjects. With only 519 subjects, it was not possible to subdivide all of the diagnoses by sex before calculating the CANDISC item-loadings. That this would have been desirable is illustrated by Fig. 1; virtually every SDQ scale shows a substantial sex difference in the magnitude of response. This design problem also made it impossible to include any questions from the for men only or for women only sections of the SDQ. A goal of any future recalculation of the SDQ will be to create unique scales for males and females. It is likely that increased diagnostic specificity could be achieved. Similarly, 111 out of 175 questions of the SDQ were not used in this analysis, and many of these have test–retest reliabilities over 0.6. It will be a future goal to refine the scales using more subjects and more SDQ items.

Another limitation of the present findings resides in the control group, largely normal individuals under age 35. It would have been preferable to have an age-stratified control matched to the clinical patients. There are clinical data to suggest that some sleep disorders worsen with age, which makes age-norming a future goal of research with the SDQ. Also, because only about 20% of controls had NPSGs, one cannot be sure that the rest were entirely free of sleep disorder. Because controls were not screened for psychiatric disorder, it is possible that some had such conditions, which would confound the calculation of the PSY scale. Efforts will be made to acquire more carefully screened, polysomnographically normal age-matched controls in future.

The multicenter design of this study may have introduced some diagnostic errors into the results. Because a given center provided only one diagnostic group, it is possible that there was variation in differential diagnosis or implementation of diagnostic criteria between centers. A better design would involve a larger number of cases, all from one center. A further refinement would be to make diagnoses via a structured interview using two or more clinicians simultaneously, so that inter-rater reliability of diagnosis could be assessed.

What role should the SDQ play in screening individuals for the possibility of sleep disorder at this early stage of its development? The whole SDQ is too long

TABLE 5. *MANOVA[a] and one-way ANOVAs of scale scores*

Dependent variable	Source	df	Sum of squares	Mean square	*F* value	p	Significant pairwise comparisons (Bonferroni at alpha = 0.05)
SA	Group	4	38,091	9,522	163.20	<0.0001	A-C, A-P, A-N, A-M
NAR	Group	4	11,508	2,877	31.49	<0.0001	N-C, N-P, N-M, N-A
PSY	Group	4	6,952	1,738	35.11	<0.0001	P-C, P-N, P-M, P-A
PLM	Group	4	3,107	776	18.48	<0.0001	M-C, C-P, C-N, C-A

This multivariate analysis of variance was a post hoc confirmatory test of the final SDQ scales after item deletion using Cronbach's Alpha criteria. Groups in the pairwise comparisons were: A = Sleep Apnea group; C = Control group; M = Nocturnal Myoclonus/PLMD group; N = Narcolepsy group; P = Psychiatric group.

[a] MANOVA significance: Wilks' Lambda = 0.249, $F = 56.02$, df = 16/1,558, $p < 0.0001$; Pillai's Trace = 1.005, $F = 43.04$, df = 16/2,052, $p \leq 0.0001$.

to be used as a general-population epidemiological screening tool, and base rates would be lower than those observed in this sample, so more false positives would be detected. In the future, 5–10 items will be selected from the SDQ and studied for the purpose of general-population screening.

The usefulness of SDQ in supporting a diagnosis in patients suspected by a general practice physician of having sleep apnea or narcolepsy seems to be confirmed by the present study. Although statistically significant results were also found for psychiatric and PLMD groups, the PPVs of these scales were low. We suggest that SDQ be used as a confirmatory diagnostic tool after the clinical interview in general practice, in much the same way that psychological testing is used to support suspected psychiatric diagnoses. Although we do not show the supporting data here, we have also used it to decide whether to schedule a patient for NPSG alone or NPSG plus MSLT, based upon scores on the SA and NAR scales. In those laboratories doing limited or ambulatory polysomnography, such information might be all the more important.

Although it would be feasible to give subjects a shortened questionnaire consisting only of the items that appear on the four scales, we recommend that the whole 175-item SDQ be given. Testing time (under 30 minutes) is not much longer, and it is anticipated that scoring scales for other disorders will be derived from this pool of questions as usage increases. Shortened questionnaires will not allow calculation of the new scales.

The fact that sleep disorders have overlapping symptoms to some degree is suggested by the high NAR scale in the PSYCH group—perhaps related to hypnagogic hallucination in depressives and schizophrenics with very short rapid eye movement sleep latencies. Likewise, there are some NARCO and MYOCL group patients who have fairly high scores on the SA scale. In the companion article (SDQ II), it can be seen that this represents the accurate detection of dual pathology by SDQ.

A final disclaimer: the scales presented are not a replacement for a clinical assessment by a trained sleep clinician plus polysomnography. They should neither be used as the only diagnostic tool nor to make treatment decisions in the absence of such an assessment. The scales do allow a clinician to compare the SDQ responses of a new patient to the 519 persons in this study.

A manual about the scoring of the SDQ, scoring software and copyright release are available by mail from the first author. SDQ is currently being translated into the major European languages.

Acknowledgements: This work was supported in part by Internal Funds of the University of Michigan Department of Psychiatry and by the Stanford Sleep Disorders Foundation, Palo Alto, CA.

Dr. German Nino-Murcia died in July 1993. We will always remember his quiet, helpful contributions to this project. —A.B.D.

REFERENCES

1. Kapuniai LE, Andrew DJ, Crowell DH, Pearce JW. Identifying sleep apnea from self-reports. *Sleep* 1988;11:430–6.
2. Leigh TJ, Bird HA, Hindmarch I, Constable PDL, Wright V. Factor analysis of the St. Mary's Hospital Sleep Questionnaire. *Sleep* 1988;11:448–53.
3. Buysse DJ, Reynolds CF, Monk TH, Berman SR, Kupfer DJ. The Pittsburgh Sleep Quality Index: new instrument for psychiatric practice and research. *Psychiatry Res* 1989;28:193–213.
4. Webb WB, Bonnet M, Blume G. A post-sleep inventory. *Percept Mot Skills* 1976;43:987–93.
5. Domino G, Blair G, Bridges A. Subjective assessment of sleep by Sleep Questionnaire. *Percept Mot Skills* 1984;59:163–70.
6. Hoddes E, Zarcone V, Smythe H, Phillips R, Dement WC. Quantification of sleepiness: a new approach. *Psychophysiology* 1973;10:431–6.
7. Parrott AC, Hindmarch I. The Leeds sleep evaluation questionnaire in psychopharmacological investigations—a review. *Psychopharmacology* 1980;71:173–9.
8. Horne JA, Ostberg O. A self-assessment questionnaire to determine morningness-eveningness in human circadian rhythms. *Int J Chronobiology* 1979;22:79–91.
9. Douglass AB, Carskadon MA, Houser R. Historical data base, questionnaires, sleep and life cycle diaries. In: Miles LE, Broughton RL, eds. *Medical monitoring in the home and work environment.* New York: Raven Press, 1990:17–27.
10. Guilleminault C, ed. *Sleeping and waking disorders: indications and techniques.* Menlo Park: Addison-Wesley Publishing Co., 1982.
11. Douglass AB, Bornstein R, Nino-Murcia G, Keenan S. Creation of the "ASDC Sleep Disorders Questionnaire". *Sleep Res* 1986; 15:117.
12. Douglass AB, Bornstein R, Nino-Murcia G, et al. Test–retest reliability of the Sleep Disorders Questionnaire (SDQ). *Sleep Res* 1990;19:215.
13. Association of Sleep Disorders Centers Classification Committee. *Diagnostic classification of sleep and arousal disorders,* 1st ed. New York: Raven Press, 1979.
14. Endicott J, Spitzer RL. A diagnostic interview: the Schedule for Affective Disorders and Schizophrenia. *Arch Gen Psychiatry* 1978;35:837–45.
15. Carskadon MA, Dement WC, Mitler MM, Roth T, Westbrook PR, Keenan S. Guidelines for the multiple sleep latency test (MSLT): a standard measure of sleepiness. *Sleep* 1986;9:519–24.
16. Royston JP. An extension of Shapiro and Wilk's W test for normality to large samples. *Applied Statistics* 1982;31:115–24.
17. Fletcher RH, Fletcher SW, Wagner EH. *Clinical epidemiology, the essentials,* 2nd ed. Baltimore: Williams and Wilkins, 1988.
18. Briggs SR, Cheeks JM. The role of factor analysis in the development and evaluation of personality scales. *J Pers* 1986;54: 106–48.