Sleep Science and Practice

RESEARCH

Open Access

Comparative validation of the Sleep Disordered Breathing scale of the Sleep Disorders Questionnaire 2nd edition and four other sleep apnea questionnaires against polysomnography

Alan Douglass[1,5,6*], Dianne Groll[3], Kathleen Biard[2], Caitlin Higginson[1], Eva Savage[4], Karina Fonseca[2], Malika Lanthier[2], Defne Oksit[2], Chloé Léveillé[2], Arianne Côté[2], Noémie Leblanc[2] and Rébecca Robillard[1,2]

Abstract

Background The original Sleep Disorders Questionnaire (SDQ, 1994) contained a 12-item sleep apnea subscale (SDQ-SA). SDQ was shortened by factor analysis in 2024 (SDQ-2) and contains a 14-item subscale "Sleep Disordered Breathing" (SDQ2-SDB). This is the first external validation study of the new scale using polysomnography, with a concurrent validation against four commonly used sleep-disordered breathing questionnaires.

Methods Members of a large urban police force (n = 733) completed the SDQ2-SDB, Berlin, STOP-Bang, NoSAS, and GOAL questionnaires. A subgroup (n = 101) also volunteered for a laboratory polysomnogram (PSG). SDQ2-SDB scores were analyzed by ANOVA in both samples. For the NPSG subgroup, Receiver Operating Characteristics (ROC) of the five apnea scales were calculated at a criterion of AHI ≥ 15. A correlation matrix among the five scales was also calculated.

Results The NPSG subgroup with higher measured AHI showed significantly higher SDQ2-SDB scores. In the whole sample, SDQ2-SDB scores were significantly higher in males compared with females, in older compared to younger individuals, and in those with higher BMI. Age x Sex, Age x BMI and BMI x Sex effects were also significant. The areas under the ROC curve (AUC) were: SDQ2-SDB = 70%, Berlin = 65%, STOP-Bang = 73%, GOAL = 81%, NoSAS = 85%. The NoSAS AUC was significantly higher ($p < 0.03$) than those of the SDQ2-SDB, STOP-Bang, and Berlin scales but the latter three scales did not differ significantly from each other. Positive (PPV) and negative (NPV) predictive values of the SDQ2-SDB were 46% and 84% respectively, with the other scales ranging between 35 to 52% PPV and 74 to 89% NPV. The SDQ2-SDB correlated with the other four questionnaires with rho ranging from 0.558 to 0.855.

Conclusions The SDQ2-SDB scale demonstrates adequate validity in a non-clinical sample. Its performance is similar to questionnaires commonly used for sleep apnea, like the Berlin and STOP-Bang scales.

*Correspondence:
Alan Douglass
alan.douglass@theroyal.ca

Full list of author information is available at the end of the article

Keywords Sleep apnea, Questionnaires, External validation, Concurrent validation, Polysomnogram, Diagnostic screening, Non-clinical participants, ROC methodology

Brief summary

Study rationale

This was an external validation of the 14-item Sleep-Disordered Breathing (SDB) scale of the Sleep Disorders Questionnaire, second edition (SDQ-2, published 2024) using polysomnography in a non-clinical sample. SDQ2-SDB was also compared with other apnea questionnaires – Berlin, STOP-Bang, NoSAS, and GOAL – for concurrent validation.

Study impact

SDQ2-SDB scores can be derived from existing SDQ (1994) data or directly from the SDQ-2. The SDQ2-SDB is self-reported, requiring no staff input and has been translated into several languages. There was a trend for certain SDQ2-SDB items to show differences between males and females. This study broadens the choice of validated sleep apnea questionnaires.

Background

In the 50 years since Guilleminault and Dement published their foundational papers (Guilleminault et al. 1976a, b), much has been learned about its diagnosis and treatment, as well as that of milder forms of sleep-disordered breathing (SDB). The international prevalence of all forms of SDB has recently been estimated to be about 12% (Benjafield et al. 2019). While the substantial long-term consequences of untreated SDB for health and public safety have been documented (Dodds et al. 2020), management of the disorder has notably been hampered by inconsistent case-finding and relative paucity of diagnostic facilities (Dodds et al. 2020; Miller and Berger 2016).

Today, the nocturnal polysomnogram (NPSG) is the established gold-standard for SDB diagnosis, whether it be a laboratory study or an ambulatory multi-channel recording. However, the cost of these recording methods and the specialized medical expertise required to interpret them is quite substantial, easily exceeding 3,000 USD in western countries (Natsky et al. 2021). While such facilities are reasonably well-dispersed in North America and Europe, the same cannot be said for other parts of the world, for rural areas, or for disadvantaged parts of large cities. This suggests the need for more accessible screening tools to identify potential SDB cases. Three major methods of detecting SDB have been suggested in the literature.

First, physicians could simply identify cases in their practice that match published diagnostic criteria, such as those of the American Academy of Sleep Medicine. However, this method has proven to perform poorly at detecting SDB when compared to common sleep apnea questionnaires (Duarte et al. 2023).

Second, physicians could fill out rating scales or measurements of anatomic features that they observe during the physical exam, such as soft palate morphology (Yu and Rosen 2020), tonsil size (Friedman et al. 1999), body mass index (Takegami et al. 2009), upper airway breath sounds (Douglass and Kaluzienski 2024), neck circumference, or waist-hip ratio (Xu et al. 2020). Only some of these methods have been fully validated by test theory.

Third, questionnaires completed by the patient have been developed, although some are not truly based on self-report as they require one or more items to be completed by medical personnel. Approximately twenty such questionnaires have been developed over the past 35 years, four of which will be considered in the present paper: STOP-Bang (Chung et al. 2008; Shi et al. 2025), Berlin Questionnaire (Netzer et al. 1999), GOAL (Duarte et al. 2020), and NoSAS (Marti-Soler et al. 2016). Their relative performance has been compared in several reviews and meta-analyses (Abrishami et al. 2010; Fedson et al. 2012; Lomeli et al. 2007; Kim et al. 2015; Casale et al. 2008; Xing et al. 2023; Wang et al. 2021; Chen et al. 2025; Gauld et al. 2024; Gauld et al. 2023).

It is notable that most of the above screening questionnaires have been derived and tested in persons already suspected of SDB and referred to a sleep disorders clinic; however, the prevalence of OSA in such a population is much higher than in the general population. This was our rationale for studying a non-clinical sample.

The original 176-item Sleep Disorders Questionnaire (Douglass et al. 1994) was developed by canonical discriminant function analysis (CDFA) of clinical materials gathered at Stanford University Sleep Disorders Clinic. While it was never factor analyzed by the original authors, another study (Schwegler et al. 2006), using only 45 of the 176 SDQ items, factor analyzed SDQ responses of 4901 members of the general public and identified a sleep apnea scale that was very similar to the SDQ-SA scale (also known as SA-SDQ) that had been found by CDFA in the original SDQ paper. Over the years, SDQ-SA has been employed to screen various sub-specialty clinical populations: epilepsy (Weatherwax et al. 2003), asthma (Teodorescu et al. 2009), COPD (Valipour et al. 2011), stroke survivors (Bassetti et al. 1996), hemodialysis (Lee et al. 2009), and diabetes (Seetho and Wilding 2013). It has also been used routinely in numerous sleep disorder clinics and translated into several major languages (Vergara-Aguilar et al. 2023; Piperidou et al. 2008;

Methipisit et al. 2016; Valipour et al. 2007; Alnesr et al. 2024; Hersberger et al. 2006).

The first factor analysis of the full 176-item SDQ, involving 2131 persons, resulted in a questionnaire of 66 items named "Sleep Disorders Questionnaire Second Edition (SDQ-2)" (Biard et al. 2024). This factor analysis found seven clinical scales, one of which was a 14-item Sleep-Disordered Breathing scale (SDQ2-SDB) that proved to include 10 of the 12 items from the original SDQ-SA scale. The present report is the first validation study of the SDQ2-SDB scale.

The objective of the present study was to validate the SDQ2-SDB scale against NPSGs and four existing sleep apnea questionnaires in a group of healthy individuals, both male and female, who were not selected because of suspicion of sleep apnea. This led us to approach the employee association of a large urban police department, which agreed to disseminate the 66-item SDQ-2 to its members. Specifically, we aimed to find a valid cut-off score for the SDQ2-SDB based on NPSG validation, then apply it to a larger sample serving as a proxy of the general population in order to estimate the number of cases of SDB.

Methods

All active and retired members of the police force of a major Canadian city were provided with a joint communication from the Police Service, Police Association and Senior Officer Association. In order to encourage participation while maintaining confidentiality, members were asked to reply to the survey (hosted at a university on *Qualtrics*, Provo, UT, USA. https://www.qualtrics.com) from their personal email, since their employer had ownership of the IT infrastructure at their workplace.

Due to the anonymity of the online full sample (n=733), it was not a random sample of the members. After completing the online questionnaires, a subgroup responded to a random invitation to have an in-lab polysomnogram ("NPSG subgroup", n=101), for which they received monetary compensation. It was only after the NPSG that participants could be classified as normal/mild (AHI < 15) or moderate-to-severe SDB (AHI ≥ 15). This resulted in Table 1. Although questionnaire-based data collection began in early 2019, the sleep laboratory was closed because of the COVID epidemic until autumn 2022, after which studies resumed until the end of the project in December 2023.

Data collection

The first page of the online questionnaires displayed the information and consent form approved by the REB. The server then provided an anonymous randomly-generated identification (ID) number to anyone accessing the questionnaires. No compensation was offered for completing them. Single-session complete replies were the norm.

Questionnaires included the 66-item SDQ-2, 14 items of which comprise the SDQ2-SDB scale. This scale includes a pair of 5-level categorical items (Q59 height, Q60 weight) that permit the estimation of BMI via a published algorithm (Biard et al. 2024). In the NPSG subgroup, this estimated BMI was compared to the BMI that was physically measured that night, to confirm the validity of the algorithm. In addition, collar size of the dress uniform was recorded in binary fashion (+/- 41 cm for females, +/- 43 for males) then used along with the self-reported height (Q59) to produce a "neck-height ratio".

Other questionnaires included the 10-item Berlin Questionnaire (Berlin-Q), and the 8-item STOP-Bang. In addition, participants were asked to provide the following data: work status (active or retired); job type (civilian or uniformed); age, sex, height, and weight to facilitate completion of the NoSAS and GOAL questionnaires.

At the sleep laboratory, in-person anthropometric data were collected: height, weight, BMI, and waist-hip ratio. Prior to retiring for the night, participants completed the SDQ-2, Berlin and STOP-Bang questionnaires for a second time to ensure temporal alignment between symptom reports and NPSG data. The NoSAS and GOAL were completed based on measured values.

Sleep studies

NPSGs were performed in individual bedrooms. The EEG electrode placement conformed to the international 10–20 system. The set-up of nasal/oral airflow, chest expansion, oximetry, EKG, and EMG (chin and legs) was based on the American Academy of Sleep Medicine (AASM) guidelines (Berry et al. 2015).

Data were recorded using the Embla N-7000 system and the RemLogic software (version 3.4.4, Natus Medical Incorporated, WI, USA). To reduce error variance, all NPSGs were scored by the same Registered Polysomnographic Technologist (RPsgT) using AASM scoring criteria. NPSGs were then audited by the first author, a board-certified sleep specialist physician.

Since some police members worked rotating shifts, their laboratory bedtime was adjusted to match their regular bedtime on their current shift. This involved starting some PSGs in the daytime, which were conducted in a windowless sound-proofed laboratory bedroom.

Statistical analysis

The GOAL, NoSAS, STOP-Bang and Berlin-Q were scored according to established methods. The Berlin-Q has 10 questions, which are divided into three categories that are in turn scored as "1, positive" or "0, negative", resulting in an overall score of 0—3. If 2 or

Table 1 NPSG and anthropometric data of the NPSG subgroup (n=101)

	Males, n=42 AHI < 15/h			Males, n=27 Mod.-Severe SDB		Females, n=29 AHI < 15/h			Females, n=3 Mod.-Severe SDB		
Variable	**Mean**	**SD**	***Within-sex sig.***	**MEAN**	**SD**	**Mean**	**SD**	***Within-sex sig.***	**MEAN**	**SD**	***Between-sex SDB signif.***
Age @	36.9	10.5	**p=0.002***	46.1	10.6	41.4	12.2	N.S.	47.2	13.3	N.S.
Collar size (counts)#	6	--	N.S.	20	--	2	--	N.S.	3	--	N.S.#
Hgt (cm)	176.9	8.3		179.8	7.3	170.9	5.7		164.7	11.5	
Wgt (kg)	85.1	12.9		99.5	15.9	74.5	11.5		79.4	15.0	
BMI_meas (kg/m^2) @	27.3	3.4	**p=0.0009***	30.8	4.3	25.5	3.8	**p=0.022***	29.3	4.4	N.S.
BMI_mtrx (kg/m^2) @	27.0	3.5	**p=0.0052***	30.0	4.2	25.5	3.9	N.S.	27.5	4.7	N.S.
Waist_Hip_Ratio @	0.942	0.043	**p=0.0007***	0.987	0.053	0.871	0.069	N.S.	0.898	0.056	**p=0.0001 ***
STOP-Bang	2.8	1.1	N.S.	3.6	1.6	2.1	1.3	N.S.	2.5	1.2	**p=0.042***
BERLIN_categories	1.0	0.9	*p=0.097*	1.4	0.8	1.2	1.0	N.S.	1.4	0.7	N.S.
BERLIN_total points	2.8	2.2	*p=0.074*	4.0	2.4	3.3	2.6	N.S.	4.2	1.4	N.S.
GOAL	1.7	0.8	**p=0.011***	2.3	0.9	0.8	0.7	**p=0.018***	1.4	0.7	**p=0.003***
NoSAS	7.2	3.5	**p=0.015***	9.6	3.6	3.4	2.9	**p=0.016***	6.8	3.8	**p=0.049***
SDQ2-SDB score	19.1	8.2	**p=0.014***	25.9	7.9	22.6	8.1	N.S.	28.0	6.1	N.S.
Sleep Period (min.)	423.7	68.2		446.4	79.5	462.2	82.7		482.0	63.4	
WASO (min.)	44.6	29.7	*p=0.066*	64.8	43.7	38.6	27.9	**p=0.021***	72.5	49.7	N.S.
Total Sleep Time (min.)	374.1	75.9		381.6	75.3	421.0	80.5		409.5	82.8	
Sleep Onset Latency (min.)	14.6	20.0		16.3	19.5	22.6	29.1		37.6	48.8	
Sleep Efficiency%	86.5	9.2		82.3	11.7	87.2	9.3		79.4	12.5	
REM Latency (min.)	93.4	46.9		106.0	57.2	115.0	59.1		144.1	71.8	
N1 sleep (minutes)	30.5	14.1	p=0.010*	51.1	39.7	26.9	13.3	*p=0.052*	43.5	22.9	N.S.
N2 sleep (minutes)	241.5	56.9		226.2	54.0	244.1	50.6		236.8	63.2	
N3 sleep (minutes)	43.0	26.0	**p=0.034***	29.1	27.1	58.3	36.0	N.S.	60.1	40.8	**p=0.009***
REM sleep (minutes)	64.2	31.8		75.2	30.7	91.7	41.5		69.0	29.3	N.S.
Number of Wakes	25.4	8.4	N.S.	28.1	12.9	19.0	6.8	**p=0.025***	27.3	12.2	N.S
WAKE minutes	59.1	46.6	**P=0.044***	81.2	54.3	61.2	53.7	**p=0.019***	110.1	70.0	N.S.
AHI (#/hr)	2.8	1.5	**p<0.0001***	24.9	21.7	1.2	1.1	**p<0.0001***	21.0	7.6	N.S.
HI (#/hr)	2.5	1.4	**p<0.0001***	20.7	15.7	1.1	1.0	**p<0.0001***	19.0	7.4	N.S
AI (#/hr)	0.23	0.39	**p=0.0005***	4.2	8.2	0.13	0.44	**p=0.0004***	1.9	1.1	N.S
OxyDesatIndex (#/hr)	2.0	1.6	**p<0.0001***	21.9	20.6	1.0	1.1	**p<0.0001***	9.7	7.2	**p=0.042***
RERA index	2.5	2.0	N.S.	2.6	3.2	1.7	1.2	**p=0.007***	3.6	6.3	**p=0.012***
RDI (#/hr)	5.2	2.6	**p<0.0001***	27.5	21.1	2.9	1.9	**p<0.0001***	18.4	8.7	N.S.
RDI supine (#/hr)	3.9	3.2	**p<0.0001***	37.1	32.7	1.7	2.2	**p<0.0001***	19.2	15.3	N.S.
RDI non-supine (#/hr)	1.8	1.4	**p<0.0001***	17.5	21.6	0.7	0.8	**p=0.002***	7.0	7.8	*p=0.068*
RDI (#/hr of REM)	6.5	5.3	**p<0.0001***	32.9	25.2	4.1	3.9	**p<0.0001***	27.4	19.3	N.S.
RDI (#/hr of NREM)	1.9	1.3	**p<0.0001***	22.7	22.2	0.4	0.4	**p<0.0001***	8.4	6.2	**p=0.009***
Apnea Arousal Index (#/hr)	0.1	0.1	**p=0.001***	2.4	5.6	0.0	0.1	**p=0.008***	0.4	0.6	N.S.
Hypopnea Arousal Index /hr	1.5	1.0	**p<0.0001***	9.6	10.0	0.6	0.6	**p<0.0001***	5.5	2.8	N.S.
Total Arousal Index (#/hr)	8.2	9.8	**p<0.0001***	21.2	17.7	6.4	6.7	**p=0.002***	17.7	12.1	N.S.
RespEvents >90% oxy sat	9.2	6.4	**p<0.0001***	77.8	59.2	4.5	5.3	**p<0.0001***	51.9	36.6	N.S.
RespEvents 81-90% sat	0.6	1.4	**p<0.0001***	49.4	96.8	0.1	0.2	**P=0.001***	6.8	11.9	**p=0.049***
RespEvents 71-80% sat	0.0	0.0	N.S.	2.7	15.3	0.0	0.0	N.S.	0.3	0.9	N.S.
Mean Oxy Sat%	96.5	1.0	**p<0.0001***	94.7	1.8	97.1	0.9	**p=0.001***	95.5	1.2	N.S.
Lowest Oxy Sat%	91.9	2.6	**p<0.0001***	86.4	5.4	92.9	2.3	**p<0.0001***	88.5	4.6	N.S.
Average Desat%	3.3	1.0	**p<0.0001***	4.2	1.0	2.3	1.6	**p=0.002***	3.7	0.7	N.S.
Average Awake Sat%	96.8	1.0	**p=0.0004***	95.4	1.5	97.6	1.0	**p=0.008***	96.6	1.0	**p=0.012***
Heart Rate mean	58.1	9.6		60.9	10.5	65.1	6.5		67.0	10.7	
Heart Rate min	26.2	2.5		26.7	4.2	26.0	1.8		25.6	0.7	
Heart Rate max	140.3	18.5		142.4	16.2	143.4	15.7		138.6	14.9	
PLMs	23.6	57.6		40.2	74.9	31.4	56.6		39.1	62.9	
PLM arousal index (#/hr)	4.2	9.2		6.6	12.2	4.1	7.2		5.4	8.4	N.S.

Data from the NPSG subgroup (n = 101) who volunteered from the full sample (n = 733) for an in-laboratory NPSG, i.e., not randomly or systematically selected. Boxed area indicates questionnaire responses. Not all comparisons were tested statistically. Tests of significant difference between independent groups employed the Wilcoxon rank sum test unless data passed a test of normality (Kolmogorov-Smirnov-Lilliefors), in which case the Welch independent t-test was used (denoted with "@"). *Within-sex sig.*: significance tests of a variable between the "Mild-Normal" group (AHI < 15/h) and the "SDB" group (AHI ≥ 15/h) of the same sex. *Between-sex SDB sig.*: comparisons of the SDB Males group with the SDB Females group. Significant uncorrected (raw) *p*-values are shown in bold type. The Bonferroni-Holm correction for multiple comparisons was then applied to the observed *p*-values to achieve a Family-Wise Error Rate of alpha=0.05; those comparisons remaining significant at p<0.05 are shown by light shading; p<0.01 by dark shading. "#" indicates number of persons with collar size > 43 cm for males, > 41 cm females, which was evaluated by chi-squared test, df = 1

more categories were positive, the Berlin-Q makes a determination of "high risk" versus "low risk" of sleep apnea. The NPSG subgroup was further divided into two groups on this basis, and their AHIs compared using the Wilcoxon-Mann-Whitney rank-sum test (R Statistical System, version 4.3.3, R Core Team, 2025). For comparison, we devised a novel way of scoring the questionnaire that involved summing all the "points" of the ten questions for a total score that ranged from 0–10. However, we report below significant ROC differences between the Berlin-Q and other scales using only the original scoring method.

To assess whether the SDQ2-SDB data met the normality assumption of parametric tests, its data were tested by the Kolmogorov-Smirnov-Lilliefors test, although it is known to be overly-sensitive when used on large integer vectors composed of summed ordinal Likert scale items like SDQ2-SDB. This is in part because such data have finite upper and lower bounds and inevitable "ties"of scores.

Prior to running analyses of variance (ANOVAs), age and BMI variables were first sorted (ascending) then divided into three levels of nearly equal size in order to fill the factor levels evenly. This resulted in BMI factor levels of < 28, 28 ≤ 30, and >30 kg/m^2 and age factor levels <40, 40 to 50, and > 50 years. In police members, BMI < 28 can be regarded as normal, rather than the usual BMI < 24, because they are required to have higher levels of physical fitness than the general public as a condition of their employment. The sex factor was not evenly balanced, since the n=101 sample contained 32 females and 69 males.

Two ANOVAs were performed using the dependent variable "SDQ2-SDB scale score" via the "*aov*" command (R Statistical System, version 4.3.3). The first was a 3-way ANOVA on the NPSG subgroup, n = 101, with main factors as defined above: age (three levels), sex at birth, and measured BMI (three levels). Due to there being no cases of females with severe apnea, AHI was not included in this ANOVA, but three subsequent 1-way ANOVAs of SDQ2_SDB were done using AHI or RDI as the factor (four levels). Males and females were analyzed separately.

SDQ2-SDB questionnaire data in the NPSG subgroup were then analyzed using modern test theory (Fabricant 2024; Wong and Lim 2011), which associates questionnaire scores with "gold standard" data (the NPSG) by means of a Receiver Operating Characteristics (ROC) analysis. This was done using R's "pROC" package (Robin et al., https://CRAN.R-project.org/package=pROC). The ROC produced a questionnaire "cutoff score" corresponding to the AHI threshold (AHI = 15/h) that was chosen to separate mild/normal individuals from those diagnosed with SDB. It also calculated the sensitivity, specificity, PPV, and NPV of the test, plus the area under the curve (AUC). ROC statistics were also calculated for each of the other four sleep apnea scales: Berlin-Q, STOP-Bang, GOAL, and NoSAS. In each case, the ROC was calculated at thresholds of 5, 10, 15, and 30 AHI; however, only data from the AHI ≥15 analysis will be reported as it gave the best area-under-the-curve (AUC) performance in all five questionnaires.

A second ANOVA of similar design was then calculated on the SDQ2-SDB results of the full sample, n = 733. The main factors were the same as the previous ANOVA: age (three levels), sex at birth, and BMI (three levels, but this time estimated using the published SDQ-2 algorithm) (Biard et al. 2024). This larger ANOVA allowed for the inclusion of two-way interaction terms.

Results

PSG subgroup, n=101

These were self-selected participants, of whom eight were working rotating shifts and requested to have their NPSG starting time adjusted to match their current shift.

S1_Fig. shows their BMIs obtained by physical measurement at the sleep lab, regressed onto their BMI calculated by algorithm from the two subjective questions on height and weight from the SDQ-2. In the observed range of BMI, the two methods gave very similar results, R^2 = 0.897, validating the algorithm. The collar-size:height ratio was highly significant by chi-squared analysis (two levels of collar size by five levels of height, χ^2 = 121.1, df = 8, p < 0.0001). Its observed and expected frequencies are shown in S2_Table. The correlation of age with SDQ2-SDB total score was r = 0.199.

Table 1 shows NPSG, questionnaire, and anthropometric values of the NPSG subgroup. p-values of univariate tests for each variable are shown in two ways: first as plain text without correction for multiple comparisons, then with asterisks to indicate significance after correction by the Bonferroni-Holm procedure (R statistical system: *p.adjust, method="holm"*). A raw p-value = 0.002 was required to remain significant at p<0.05 after B-H correction.

The NPSG subgroup was divided by sex and further divided into those with sleep-disordered breathing at AHI ≥ 15 ("SDB") and those below this level ("normal/mild"). As expected, this resulted in highly significant differences in the respiratory variables (AHI, HI, RDI, ODI, RERA index), as shown in the column titled *within-sex differences*. However, when the male SDB category was compared with the female SDB category (*between-sex differences*), females' respiratory abnormalities were observed to be fewer than males'. This was not the case with the RERA index, which was higher in females than in males. Females also had significantly lower RDI during NREM sleep and significantly more N3 sleep. No significant sex differences in RDI were observed during REM

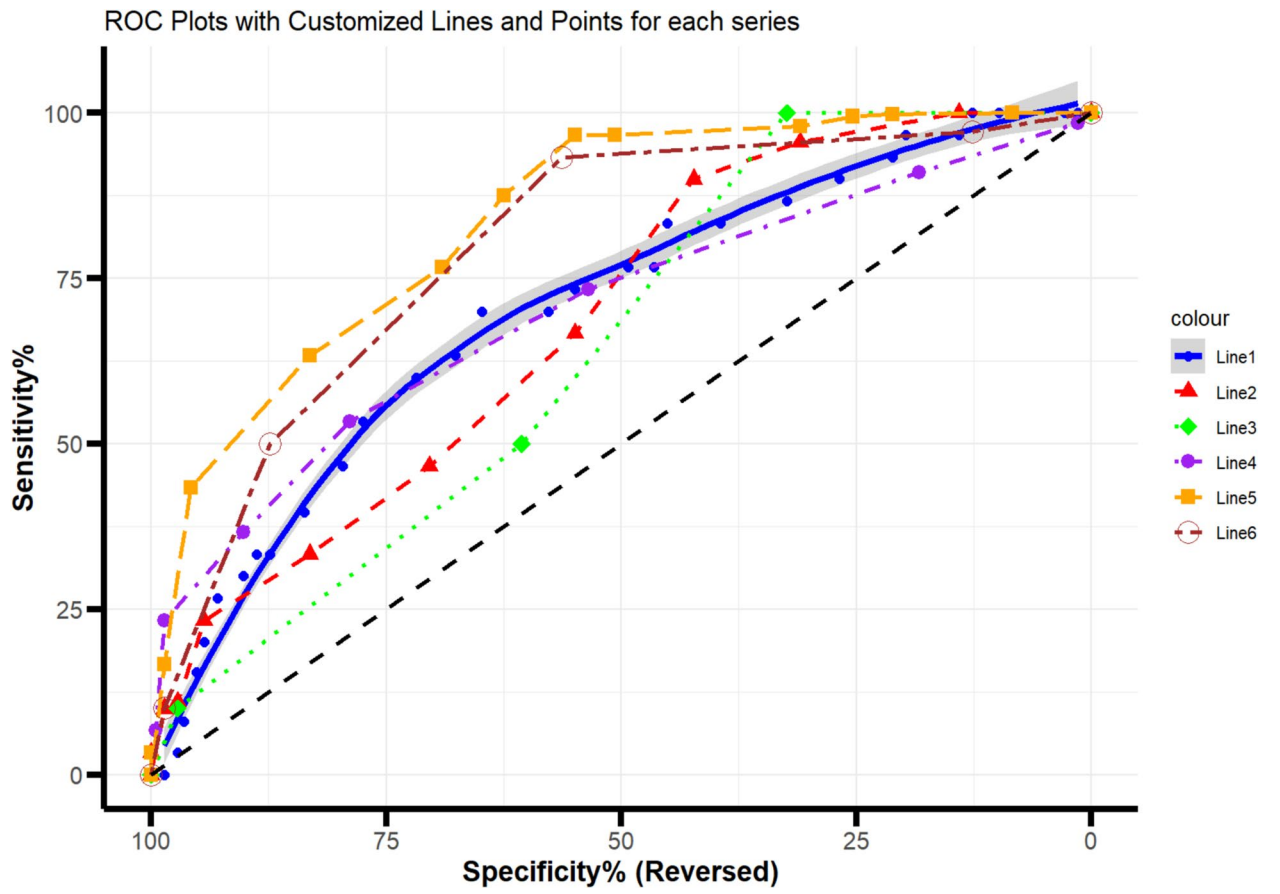

Fig. 1 ROC plots of the five sleep apnea questionnaires Line 1 = SDQ2-SDB; Line 2 = Berlin-Q #1; Line 3 = Berlin-Q #2; Line 4 = STOP-Bang; Line 5 = NoSAS; Line 6 = GOAL. All five questionnaires were assessed at a threshold of AHI > 15/h. For each, software (R-4.3.3 "pROC" package) calculated the questionnaire cut-off score that simultaneously maximized sensitivity, specificity, and area-under-the-curve (AUC). Dotted black line shows AUC = 50% (random guessing), while an AUC of at least 70% is conventionally required for clinical utility. Berlin questionnaire was plotted twice, once as a 0 – 3 "positive category count" (Berlin-Q #2, as proposed by the original authors) and again as an experimental 0 – 10 scale, summing the "points" that occur in all three of the categories (Berlin-Q #1). The SDQ2-SDB scale ROC line is shown as blue dots connected by a Loess-smoothed blue line; gray shading on either side represents + 1 s.e.m.

Table 2 Three-way ANOVA of SDQ2-SDB scale totals (n=101, NPSG subgroup)

Source	df	Sum of Squares	Mean Square	F-value	*p-value*
BMI	2	416	208.1	3.340	0.0396 *
Age	2	246	123.2	1.978	0.1441
Sex	1	17	17.5	0.280	0.5977
Residuals	95	5917	62.3		

Results of 3-way ANOVA of the SDQ2-SDB scale completed by 101 NPSG participants. Of these, only BMI was significantly related to the SDQ2-SDB score and no two-way interaction was significant. Contrasts between BMI factor levels, calculated using the Tukey HSD method, showed only a non-significant trend ($p < 0.065$) between the lowest and highest levels

Significance code: '*' $p < 0.05$

sleep. Females had significantly fewer respiratory events with oxygen saturations in the 81—90% range than did the males.

Regarding the scores of the five sleep apnea questionnaires, only the GOAL and the NoSAS showed both within-sex and between-sex significant differences, with females having lower scores. Within-sex, the Berlin and STOP-Bang showed only a statistical trend for higher scores in the moderate/severe AHI category, while the SDQ2-SDB showed significance here but only in males – likely due to low statistical power among females due to few severe apneas. The ROC analyses of these five questionnaires, Figure 1, provide better evidence of their relative test performance.

The 3-way ANOVA of the NPSG subgroup is shown in Table 2. ANOVA residuals were normally distributed (Shapiro-Wilk test: W=0.98, p-value=0.08), while homogeneity of variance was confirmed by a Levene test (F=0.72, p=0.76). Because there were so few moderate SDB cases among the females in this subgroup, and no severe ones, it was not possible to use AHI as a factor in this ANOVA. This resulted in a 3-factor design, Age x Sex x BMI, which showed that only BMI significantly affected the SDQ2-SDB score, although there was a non-significant trend for age, plus a non-significant sex difference.

Table 3 1-Way ANOVAs on SDQ2-SDB scale scores, by sex and respiratory disturbance

	Respiratory Disturbance Rate					
AHI, male	**< 5/h**	**5 ≤ 15/h**	**15 ≤ 30/h**	**>30/h**	***p<***	**source**
n	20	22	13	14		
mean	19.0	19.8	26.3	25.4	0.014*	1-way ANOVA
s.d.	8.2	7.9	5.8	8.2		
	low	low	high	high	0.001***	F-test
AHI, female	**< 5/h**	**5 ≤ 15/h**	**15 ≤ 30/h**	**--**		
n	20	9	3	--		
mean	21.6	23.6	28.0	--	0.432 (n.s.)	1-way ANOVA
s.d.	8.1	8.7	6.1	--		
RDI, female	**< 5/h**	**5 ≤ 15/h**	**>15/h**	**--**		
n	18	8	6	--		
mean	20.1	28.8	22.7	--	0.038*	1-way ANOVA
s.d.	7.0	8.6	7.5	--		
	low	high	n.s.		0.001***	Tukey HSD

Mean (SD), and *n* of the SDQ2-SDB scale scores in three separate 1-way ANOVAs. These were done because the paucity of moderate and severe cases of female sleep-disordered breathing made it impossible to use AHI as its own factor. Males and females were therefore assessed in separate ANOVAs. The male AHI analysis was significant. Pair-wise Welch F-tests showed a highly significant difference when (normal + mild) cases were compared to (moderate + severe) cases. The AHI analysis in the females was non-significant. However, when RDI was substituted as the grouping variable for females, a significant difference between normal and mild cases was observed. The n.s. in the ">15/h" category is attributable to insufficient observations in this cell. *Significance codes:* "*" p<0.05; "**" p<0.01; "***" p<0.001

Table 4 Test statistics of the five questionnaires, n=101 (PSG subgroup)

Scale Name	**Possible Scores**	**Cut-off Score**	**Sensitivity%, [95% C.I.]**	**Specificity%, [95% C.I.]**	**AUC%, [95% C.I.]**	**PPV (%)**	**NPV (%)**
SDQ2-SDB	0 - 56	23.5	70 [46.1–83.9]	65 [37.3–81.7]	70.0 [58.8–81.0]	45.8	84.0
Berlin-Q (categories)	0 - 3	1.5	50 [30.9–73.5]	61 [47.5–76.9]	64.7 [54.4–74.7]	35.1	74.0
Berlin-Q (point total)	0 - 10	3.5	67 [46.7–89.1]	55 [39.8–68.3]	69.4 [59.0–79.7]	38.6	80.0
STOP-Bang	0 - 8	2.5	73 [57.6-–90.2]	53 [38.9-–80.3]	72.7 [62.2-–83.2]	39.6	82.0
NoSAS	0 - 17	7.5	80 [59.4–94.6]	69 [61.8–91.8]	84.5 [76.0–92.1]	52.1	89.0
GOAL	0 - 4	1.5	93 [72.4–100]	56 [37.7–70.0]	80.5 [72.4–88.7]	47.0	95.0

ROC performance of the five sleep apnea scales in detection of the threshold AHI (≥15/hr), which had a prevalence of 29.7% in this sample. Male and female data pooled. *Cut-off score*: questionnaire score that simultaneously maximized sensitivity and specificity, based on the analysis in Fig. 1

PPV and NPV are highly dependent on prevalence of the condition in the population tested. Pair-wise significant differences are shown in Table 5

AUC area under the curve, *PPV* positive-predictive value, *NPV* negative-predictive value

Table 3 shows separate subsequent one-way ANOVAs by-sex of the SDQ2-SDB score, with AHI as the grouping factor for both males and females. For females, RDI was used in addition, since the ANOVA using AHI was not significant. The erratic trend in the last row of the ANOVA of female RDI (Table 3) is likely the result of few observations/low power for this comparison.

In the ROC analysis of the NPSG subgroup, at a threshold of AHI ≥ 15, all five apnea questionnaires obtained a clinically useful area under the curve (AUC), but there were significant differences between them in sensitivity, specificity, PPV, and NPV. In Fig. 1, a receiver Operating Characteristics (ROC) curve is shown for each of the five questionnaires. Four of them were plotted as segmented lines on the single graphic, whereas the fifth line – the SDQ2-SDB scale – was plotted as points connected by a line derived by Loess smoothing (Cleveland and Devlin 1988). A shaded area around the Loess line denoted ± 1 standard error.

Detailed ROC test statistics (sensitivity, specificity, AUC, NPV, PPV, and their 95% confidence intervals) are shown in Table 4 for the five apnea questionnaires. The ROC procedure maximized sensitivity and specificity simultaneously. For each questionnaire, the cut-off score that separated cases of AHI ≥15 (threshold) from those with lower AHI values is shown. For the SDQ2-SDB scale only, other AHI thresholds were also tested by ROC but not shown on graphs or tables. The AUC for AHI ≥ 5 was 60%; AUC for AHI ≥ 10 was 68%; AUC for AHI ≥ 15 was 70%; AUC for AHI ≥ 30 was 63%.

Table 5 Spearman correlations and AUC comparisons of the 5 sleep apnea scales, n=101

Scale	SDQ2-SDB	Berlin-Q	STOP-Bang	GOAL	NoSAS
SDQ2-SDB	--	0.700***	0.768***	0.715***	0.640***
Berlin-Q	0.84 (n.s.)	--	0.574**	0.558**	0.466**
STOP-Bang	0.32 (n.s.)	0.21 (n.s.)	--	0.815***	0.755***
GOAL	0.069 (trend)	0.0112*	0.065 (trend)	--	0.719***
NoSAS	0.0254*	0.0013**	0.0077**	0.26 (n.s.)	--

Upper triangle: Spearman non-parametric correlations between the apnea scales. Highest correlation was between the STOP-Bang and GOAL questionnaires; lowest was between Berlin-Q and NoSAS. *Lower triangle:* statistical significance (*p*-values) of pair-wise Delong's tests of AUCs of the five scales from Table 4. "Berlin-Q" refers to the original 3-category scoring method of this scale

Signif. Codes: '***'= p<0.001; '**'= p<0.01; '*'= p<0.05

Due to low numbers of severe female apnea cases in the NPSG subgroup, ROC cut-off scores had to be calculated by pooling male and female data. For the SDQ2-SDB scale, whose possible range is zero to 56 points, this cut-off score was 23.5. The SDQ2-SDB had the third highest values for sensitivity and specificity, exceeded only by the NoSAS and the GOAL. PPV and NPV were calculated using the observed prevalence of sleep-disordered breathing in the NPSG subgroup, which was 29.7%.

The Berlin-Q was scored in two ways: the "number of positive categories" method as proposed by the original authors, and a novel "sum of points across categories" method, published here for the first time. The latter method is provisional and investigational at this time; its validity has not yet been confirmed by other studies. Analyses in the remainder of this article employ only the original method. In the NPSG subgroup, the original method demonstrated a cut-off score of 1.5 categories, the novel method a cut-off score of 3.5 points. Although the novel method had slightly higher AUC, sensitivity, and specificity by inspection, the two methods' AUCs did not differ significantly (Delong Test, p=0.246). The original Berlin-Q publication also suggests that persons scoring positive in two or more of the three categories are at "high risk of apnea" versus "low risk of apnea" for those with lower scores. However, when the 101 members of the NPSG subgroup were grouped in this manner, there was no significant difference in average AHI between the groups: ("low" group) n=58, mean=13.01, SD=18.4; ("high" group) n=43, mean=16.1, SD= 19.3 (nonparametric Wilcoxon-Mann-Whitney rank sum test, W = 1381.5, p -value = 0.357).

Table 5 (lower portion) shows pair-wise testing of the differences between the AUCs of the five questionnaires by means of Delong's Test. The SDQ2-SDB scale's AUC was not statistically different from those of the STOP-Bang or the Berlin-Q. The NoSAS had significantly higher AUC than the SDQ2-SDB, whereas the GOAL showed only a non-significant trend in the same direction.

Table 6 Three-way ANOVA of SDQ2-SDB scale totals (n=733, website contact only)

Source	df	Sum of Squares	Mean Square	F-value	P-value
BMI	2	11919	5960	113.6	< 0.0001 ***
Age	2	6948	3474	66.2	< 0.0001 ***
Sex	1	4917	4917	93.7	< 0.0001 ***
BMI x Age	4	626	157	3.0	0.0185 *
BMI x Sex	2	504	252	4.8	0.0085 **
Age x Sex	2	595	298	5.7	0.0036 **
BMI x Age x Sex	4	67	17	0.32	0.8648 N.S.
Residuals	715	37525	52		

Questionnaire responses from an anonymous online survey of 733 members of a large urban police force. The 14-item *Sleep Disordered Breathing scale* (SDQ2-SDB) is one of seven scales of the SDQ-2. High scores on it were significantly related to higher BMI, older age, and male sex. See Fig. 2 for cell means and 95% c.i. from this ANOVA. *Abbreviations:* BMI = body mass index; Age in years; Sex (at birth) = male or female

Significance codes: '***'= p<0.001; '**'= p<0.01; '*'= p<0.05

The overall similarity of the five questionnaires was assessed by pair-wise Spearman non-parametric correlations, shown in Table 5 (upper portion). The highest correlation was between the STOP-Bang and GOAL questionnaires. Lowest correlation was between the Berlin-Q and the NoSAS. The SDQ2-SDB scale correlated above *rho* = 0.70 with all scales except for the NoSAS, with which it correlated 0.64. This pattern of correlation suggests that the concepts assessed by the five questionnaires are similar despite their differing content.

Full sample, n=733

Given the approximately 2400 members of the Police Association, this group represented a questionnaire response rate of 31%. Missing data were present. The average case-wise missing rate was 4.2%, while the variable-wise rate was 1.1%. Missing cells were filled with imputed data calculated via "multiple imputation by chained equations" ("*mice*" package in R-4.3.3). A custom "predictor matrix" selected the variables that were used to predict missing data. The method of imputation used was "classification and regression trees" (*cart*), which operates via AI techniques.

A 3-way ANOVA of this larger group, again using SDQ2-SDB as the dependent variable was performed (Table 6). Although the Kolmogorov-Smirnov test with Lilliefors correction on these data (D = 0.072, p ≤ 0.001) indicated a significant difference from normality, the over-sensitivity of this test on such data is discussed above. S3_Fig. shows a frequency histogram and Q-Q plot that revealed that this difference was mainly due to symmetrical deviation in the tails of the distribution, not excessive skew, kurtosis, or lack of centrality. The SDQ2-SDB scale was therefore accepted as appropriate for ANOVA. A Levene test confirmed homogeneity of variance in the factor levels, (F-value=1.1765, df=17/715,

p=0.2777). A graphic of residuals (S4_Table) showed them to be symmetrical about zero, while a Loess-smoothed line showed no systematic deviations or clustering at extreme scores.

The main ANOVA factors of BMI, age, and sex were highly significant. A higher SDQ2-SDB score was associated with higher BMI, male sex, and older age. All of the 2-way interactions (Age x Sex, BMI x Age, and BMI x Sex) were also significant; their factor-level means are shown in separate panels of Fig 2. Three-way interactions were not significant. To summarize these interactions, older age and higher BMI caused the significant sex differences seen in the lower categories to disappear. Average SDQ2-SDB scores differed significantly by-sex: 21.8 for males, 16.6 for females, both with a 95% c.i. under 1 score unit. Cases exceeding the cut-off score of 23.5 totaled 269, or 36.7% of respondents. Using the PPV of 0.458 from the NPSG subgroup, i.e., assuming a similar prevalence, the estimated number of actual cases of SDB in the full sample would be 269 × 0.458 = 124 individuals, or 16.9% of the group.

Sensitivity analysis

Since the full sample ANOVA showed a highly significant sex difference while the much smaller ANOVA in the NPSG subgroup showed none, this discrepancy in questionnaire response was re-examined by a type of sensitivity analysis. Returning to the NPSG subgroup, each of the 14 items of the SDQ2-SDB scale was correlated, by sex, to that person's AHI and RDI. After conversion of r to Z-score using a Fisher Z-transform to normalize the distributions, each item's correlation with AHI was then subtracted from its correlation with RDI to give a *difference score* "delta", ($r_{RDI} - r_{AHI} = delta$). Delta was then back-transformed to r and graphed in Fig. 3. Visual inspection showed that the profile of male delta scores

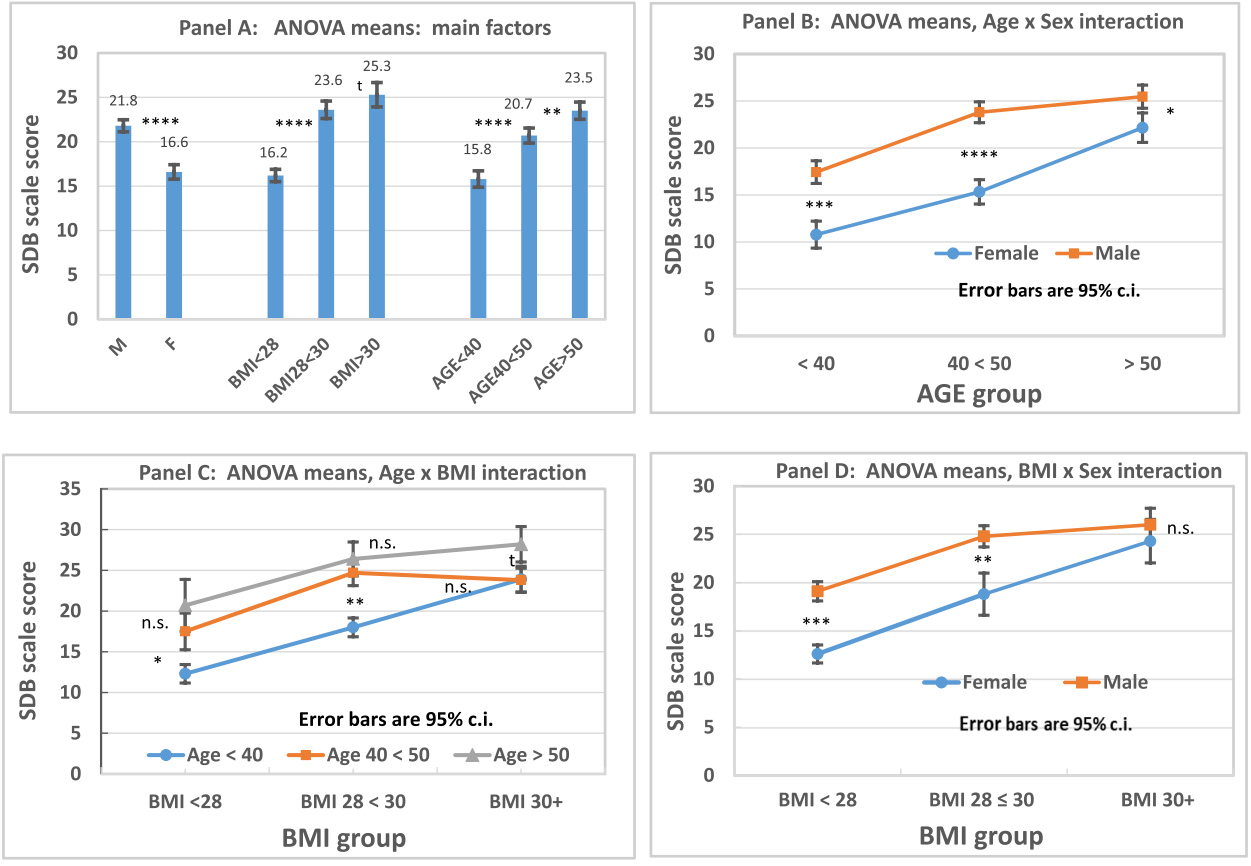

Fig. 2 SDQ2-SDB scale scores by age, sex, and BMI (n = 733, anonymous contact only)
Legend: Graphs of SDQ2-SDB mean scores (± 95% c.i.) derived from the 3-way ANOVA of Table 2. Tests of mean differences by F-test have been Bonferroni corrected. *Main factors (Panel **A**)*: sex difference was highly significant, males higher than females. There were significant differences in SDQ2-SDB scores between all age categories. SDQ2-SDB Score in the smallest BMI category was significantly lower than in the two higher categories. *2-way interactions*: *(Panel **B**)*: There were significant differences between the sexes in all age groups, more marked in the 2 younger groups. *(Panel **C**)* Youngest age group had significantly lower SDQ2-SDB scores than the 2 older age groups across both lower BMI categories but there were no significant differences in the largest BMI category. *(Panel **D**)*: Females had significantly lower SDQ2-SDB scores than males in the two smaller BMI categories, but there was no significant difference in the largest BMI category. *3-way interactions* (BMI x Age x Sex) were not significant, so were not graphed. *Significance codes:* '****' = p<0.0001; '***' = p<0.001; '**' = p<0.01; '*' = p<0.05; "t" = trend

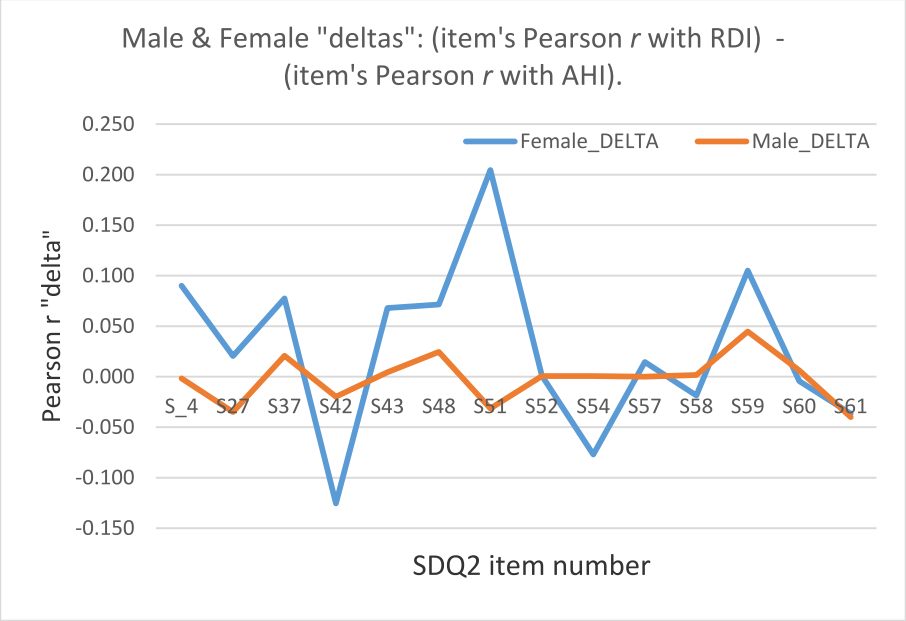

Fig. 3 Sensitivity analysis of the 14 items of the SDQ2-SDB scale
Legend: Using Pearson correlation, participants' scores on each of the 14 items of the SDQ2-SDB scale were correlated, by sex, to their AHI and RDI. SDQ-2 question items are numbered horizontally in this graph. To reveal sex differences in the pattern of response, the correlation of each item's response to AHI was subtracted from its correlation to RDI to produce a "delta" value ($r_{RDI} - r_{AHI}$ = delta). Deltas ±0.04 about zero indicated that the SDB scale predicted polysomnographic AHI and RDI equally well, which was the case in males; however, the magnitude of female deltas was significantly greater than this range ($\chi^2 = 9.81$, df = 1, p = 0.0017), shown in blue line on the graph

was close to zero for all items – i.e., their SDQ2-SDB score predicted AHI or RDI equally well – while the profile of female deltas deviated significantly from zero ($\chi^2 = 9.81$, df = 1, p = 0.0017). This suggested some possible reasons for the significant sex effect in the full sample ANOVA.

The main SDQ-2 items contributing to the different profile in females were: #4 "snore loudly and bother others", #42 "awaken gasping for breath", and #51 "presence of hypertension". Using the above formula, the calculations for each item's delta by sex were: #4 male 0.248 - 0.251 = −0.003, #4 female 0.131 - 0.041 = 0.091; #42 male 0.010–0.030 = −0.020, #42 female 0.050 - 0.175 = −0.126; #51 male 0.170 - 0.201 = −0.032, #51 female 0.183 - (−0.388) = 0.221. In other words, females more rarely endorsed loud snoring that bothered others than did males and this correlated more with their RDI than their AHI. Awakening gasping for breath was positively correlated to respiratory abnormalities in both males and females, but in females the greater correlation was to AHI, not RDI. History of hypertension was positively correlated to both RDI and AHI in males but *negatively* correlated to both of these in females – and more negatively to AHI than RDI.

Discussion

Validation of the SDQ2-SDB scale's ability to predict sleep-disordered breathing has been achieved, although the paucity of severe apneas among the females forced the pooling of male and female responses for the ROC analysis. The result was a "unisex" cut-off score of the scale to indicate the likelihood of AHI ≥ 15, rather than separate male and female cut-offs. The latter will be provided at a later date after more data have been gathered.

Given that 8 of the 101 members undergoing NPSG were on a non-daytime shift, it cannot be asserted that all circadian effects on their sleep and respiration were mitigated merely by moving their bedtime to match their current shift. As a result, our sample may have been biased in comparison with the general population. That said, some fraction of the general population also does shift work.

The SDQ2-SDB had the second highest specificity among the five questionnaires and a similar sensitivity to the STOP-Bang and Berlin-Q. In the female columns of Table 1, however, none of these three scales found a significant within-sex difference between normal/mild AHI versus moderate/severe AHI. This was almost certainly the result of too few females (three) in the moderate/severe category. For males, in contrast, this comparison was significant (p = 0.014). However, overall significance emerged when male and female responses were pooled. In contrast, the NoSAS and the GOAL found significance in both males and females separately, without pooling the sexes. This pattern is likely the result of Type-II error due to insufficient affected female participants, i.e., a power effect.

The ANOVA of the full sample demonstrated SDQ2-SDB's significant and robust increases with older age and higher BMI. There was also a significant sex difference, males more severe. These findings match known clinical features of sleep apnea and hypopnea, supporting the construct validity of the SDB scale.

Using the original SDQ-SA scale, it has been reported (Valipour et al. 2007) that women exhibit symptoms that are different from men for the same SDB diagnosis, and that this pattern may be due to symptoms of SDB being milder or different prior to menopause (Perger et al. 2019). In the present study, females had significantly more RERAs than males and a different pattern of response to the SDQ2-SDB items. While the total scale score correlated similarly to both AHI and RDI in males, certain items of the scale were more related to RDI than AHI in females. In addition, *lower* scores on the hypertension question (SDQ-2 #51, "I have high blood pressure or once had it") predicted a higher RDI in females, which has also been reported in a recent large study (Cano-Pumarega et al. 2017).

Future work should therefore expand the number of females tested with a view to clarifying differences in the male versus female SDB syndromes. While the sensitivity analysis of Fig. 3 is suggestive of a different pattern of SDQ2-SDB response in females, more detailed sensitivity analysis of the items could be done in future when there are more participants.

Similarly, it has been suggested that the sensitivity of the Berlin-Q and the NoSAS, while useful in men, is inadequate to screen women for sleep apnea (Bauters et al. 2020) and may require a different cut-off score. Sex- and age-specific cut-off scores could also be created in the future for SDQ2-SDB, as suggested by the whole group ANOVA factor levels in Fig. 2, when larger datasets become available.

There is even a suggestion (Gamaldo et al. 2018) that none of the existing sleep apnea questionnaires fully captures the pathological range of the disorder in women, so that in future a much broader approach needs to be taken. This could include questions on quality of life, psychopathology, and social factors in addition to consideration of various sub-populations. While these issues are not covered within the SDQ2-SDB scale, other scales in the full 66-item SDQ-2 questionnaire do cover some of them. Depression, anxiety, and insomnia are assessed in the InsPSY and InsSUBJ scales; substance use affecting sleep in the SUBST scale; daytime sleepiness in the EDS scale; and insomnia due to RLS/PLMD events in the InsPLM scale. In future articles, validations of these other scales will be presented.

The high correlation (Pearson r = 0.94) between the questionnaire-derived and physically-measured BMIs validates the use of self-report data from the SDQ-2 to calculate BMI rather than requiring measurements and calculation by medical personnel. This allows SDQ2-SDB to be a completely stand-alone self-report questionnaire, in contrast to the other four apnea questionnaires. However, this finding could be unique to the police group studied and might not be obtained in sicker or older individuals.

While this study population was dissimilar to patients attending a sleep disorders clinic, it was likely similar to the general population. On the other hand, the relatively high prevalence of SDB in the NPSG subgroup suggests that a proportion of participants may have self-selected based on their pre-existing suspicion of sleep apnea. In addition, the predominantly mild-to-moderate sleep apnea syndromes displayed by the police members made it hard to get simultaneously high sensitivity and specificity on ROC analysis. It will be interesting to see the performance of the SDQ2-SDB scale in patients who are waiting for a sleep disorders clinic appointment.

By inspection, the ROC performance of the SDQ2-SDB scale appears to be better than the Berlin-Q and about the same as STOP-Bang but these differences were not statistically significant. In contrast, while the GOAL and NoSAS had significantly higher sensitivities than the SDQ2-SDB scale, these occurred at specificities that were approximately the same as the SDQ2-SDB's. In addition, the GOAL and NoSAS require intervention of medical personnel to complete, whereas the SDQ2-SDB does not.

While the performance of the Berlin-Q, when scored by the novel method of "summing the points", had slightly better sensitivity and specificity than the traditional method, this difference was not statistically significant. On the other hand, the traditional "low apnea risk" versus "high apnea risk" predictions of the Berlin-Q did not show a significant difference between normal/mild and moderate/severe AHI groups in the present study. This prediction by Berlin-Q was therefore not supported by our data.

One benefit of the SDQ2-SDB being validated in a police force is that it can now be administered to other police forces and possibly emergency services or military groups as part of wellness/risk management initiatives. Early identification of members at risk of sleep-disordered breathing could facilitate better occupational functioning and public safety. To this end, the SDQ2-SDB has been released to the Police Association that formed the basis of this study. They plan to provide it confidentially to their members.

There are other limitations to the above findings. Although studying police members may produce replies that approximate the general population versus patients awaiting sleep clinic referral, this may have biased the results in the opposite direction, i.e., "supernormal." Because police members had to have achieved graduation

from at least high school and in most cases college, their reading comprehension and language fluency would likely have been higher than in some patient groups. For example, their replies to the height and weight questions might have been more accurate, which would in turn make prediction of their BMI more precise than in clinical groups. In addition, only persons who spoke English fluently were assessed in this study, in contrast to some patient groups in which there could be difficulty understanding some of the SDQ2-SDB items.

Those police members volunteering for NPSGs may have self-selected in a non-random way that increased their rate of sleep apnea compared to other members. Even if true, there were few cases of serious sleep apnea such as would be expected in clinical samples. Clinical patients' SDQ2-SDB results might show higher sensitivity and specificity than we found, or it is possible that the questionnaire would perform more poorly in severe apnea cases due to cognitive impairment or multiple comorbidities. Further validation should therefore include larger numbers of participants overall, but especially more women.

As noted above, the original SDQ-SA scale was used in a number of sub-specialty clinics dedicated to various medical and psychiatric diagnoses. Several of those authors proposed lower cut-off scores to diagnose apnea than those reported in the 1994 SDQ paper; this might also prove to be the case with SDQ2-SDB. Also, the age distribution of the police members did not include teenaged or geriatric cases. In future, therefore, it would be wise to reassess the SDQ2-SDB scale in a wide variety of specialty populations, as well as systematically accounting for age and sex. Although the unisex ROC cut-off score calculated in the NPSG sample may suffice for general population screening for the time being, this would need to be confirmed in larger samples. One goal might be to derive different cut-off scores for different subpopulations and to separate male versus female results. Another would be to see what predictions can be made of hypopnea or upper airway resistance alone, given that this study focused only on AHI.

Conclusions

This unisex validation of the SDQ2-SDB scale in a police force sample provisionally supports detection of SDB in members of the general population if the above caveats are kept in mind. Statistically, its performance in the present sample was found to be not significantly different from that of the Berlin-Q and STOP-Bang on sensitivity, specificity or AUC. Because sleep-disordered breathing is a highly prevalent yet often undiagnosed disorder, adding the SDQ2-SDB to the existing screening tools may facilitate increased detection. Since the full 66-item SDQ-2 is already translated into several major languages, and because its other scales cover areas that can be comorbid with sleep-disordered breathing (psychological symptoms, substance use, insomnia, daytime sleepiness) a more accurate screening might be achieved with use of the full 66-item SDQ-2, particularly in women. Still needed in the future will be sex- and age-specific cut-off scores and norms for the SDQ2-SDB scale. A printable copy of SDQ2-SDB is attached as S5_Doc. A look-up table showing corresponding item numbers on the SDQ2-SDB, SDQ-2, and original SDQ is shown in S6_Table.

Abbreviations

AHI	Apnea-Hypopnea Index
AUC	Area Under Curve (of an ROC analysis)
ANOVA	Analysis of Variance
BMI	Body Mass Index
CAD	Canadian dollar
CDFA	Canonical Discriminant Function Analysis
HR	Heart Rate
IRB	Institutional Review Board
NPSG, PSG	Nocturnal polysomnogram
NPV	Negative Predictive Value
ODI	Oxygen Desaturation Index
OSA	Obstructive Sleep Apnea
PSG subgroup	Participants who had an NPSG
PPV	Positive Predictive Value
RDI	Respiratory Disturbance Index
REB	Research Ethics Board (an IRB)
RERA	Respiratory Event Related Arousal
ROC	Receiver Operating Characteristics (analysis of a test)
SDQ	Original Sleep Disorders Questionnaire, 1994
SDQ-2	Sleep Disorders Questionnaire, 2nd Edition, 2024
SDQ2-SDB	Sleep Disordered Breathing subscale of SDQ-2
SDB	Sleep Disordered Breathing

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s41606-026-00194-7.

Supplementary Material 1.

Acknowledgement

We thank Matthew Skof, President of the Ottawa Police Association at the inception of this project, for his leadership in securing project approval and funding from the Association. His concern for the wellbeing of his fellow members has been an inspiration to us all

Authors' contributions

Alan Douglass 1,5,6,7 adouglass@theroyal.ca -- theory, design, statistics, writing the article. -- CORRESPONDING AUTHOR, EMAIL: alan.douglass@theroyal.ca Dianne Groll 3 grolld@queensu.ca theory, online database, writing the article Kathleen Biard 2 [kbiard@uottawa.ca](mailto:kBiard@uottawa.ca) theory, statistical analysis, writing the article Caitlin Higginson 1 Caitlin.Higginson@theroyal.ca RPsgTech, supervision & scoring of PSG recordings Eva Savage 4 evaxsavage@gmail.com recruitment, administrative Karina Fonseca 2 kfons040@uottawa.ca performed sleep studies Malika Lanthier 2 mlant045@uottawa.ca performed sleep studies Defne Oksit 2 doksi064@uottawa.ca performed sleep studies Chloé Léveillé 2 cleve050@uottawa.ca performed sleep studies Arianne Côté 2 acote104@uottawa.ca data entry & management Noémie Leblanc 2 nlebl061@uottawa.ca data entry & management Rébecca Robillard

1,2 Rebecca.Robillard@theroyal.ca theory, design, writing the article, laboratory director

Funding
This project was supported by a grant from the Ottawa Police Association, which contractually agreed with the IMHR to wave any control over data analysis, write-up, or eventual publication of the results.

Data availability
The anonymous raw data is available, upon reasonable request to A.D., the first author.

Declarations

Ethics approval and consent to participate
This prospective anonymous observational questionnaire protocol, along with its information and consent forms, was approved in 2019 by the Research Ethics Board (REB) of the University of Ottawa Institute of Mental Health Research (IMHR), approval number 2018029. The REB is an Institutional Review Board (IRB) that operates under the guidelines of the Declaration of Helsinki and the Canadian Tri-Council Policy Statement (TCPS-2, 2022). The project was conducted at the Royal Ottawa Mental Health Centre in the IMHR Sleep Research Unit.
The protocol provided for an initial online consent to complete the anonymous questionnaires. Subsequently, participants had the choice to forgo their anonymity by signing a second consent form to have an in-lab sleep study. No treatment or experimental manipulation was involved in this observational study, although the incidental finding of sleep-disordered breathing at an AHI > 5 in a participant would trigger their referral to our sleep disorders clinic.

Consent for publication
The REB (IRB) waived the requirement for consent to publish the data of the 733 anonymous persons completing the online questionnaires. For the 101 persons undergoing a sleep study, their signed consent allowed non-identifying group average data to be published.

Competing interests
Alan Douglass and Kathleen Biard, along with the University of Ottawa Institute of Mental Health Research, hold the international copyright on the SDQ-2 questionnaire and the SDQ2-SDB scale.

Author details
[1]University of Ottawa Institute of Mental Health Research, Ottawa, ON, Canada
[2]School of Psychology, Faculty of Social Sciences, University of Ottawa, Ottawa, ON, Canada
[3]Departments of Psychiatry and Psychology, Queen's University, Kingston, ON, Canada
[4]Psychology Department, Yorkville University, Fredericton, New Brunswick, Canada
[5]Royal Ottawa Mental Health Centre, Ottawa, ON, Canada
[6]Psychiatry Department, Faculty of Medicine, University of Ottawa, Ottawa, ON, Canada

Received: 30 January 2026 / Accepted: 12 May 2026
Published online: 25 May 2026

References

Abrishami A, Khajehdehi A, Chung F. A systematic review of screening questionnaires for obstructive sleep apnea. Can J Anaesth. 2010;57(5):423–38. https://doi.org/10.1007/s12630-010-9280-x.
Alnesr A, Zabad K, Al-Bundokji FK, Albardan H. Validation of the Arabic version of the SDQ-SA. Discov Med. 2024;1:24. https://doi.org/10.1007/s44337-024-00038-2.
Bassetti C, Aldrich MS, Chervin RD, Quint D. Sleep apnea in patients with transient ischemic attack and stroke: a prospective study of 59 patients. Neurology. 1996;47(5):1167–73. https://doi.org/10.1212/wnl.47.5.1167.
Bauters FA, Loof S, Hertegonne KB, Chirinos JA, De Buyzere ML, Rietzschel ER. Sex-specific sleep apnea screening questionnaires: closing the performance gap in women. Sleep Med. 2020;67:91–8. https://doi.org/10.1016/j.sleep.2019.10.023.
Benjafield AV, Ayas NT, Eastwood PR, et al. Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis. Lancet Respir Med. 2019;7(8):687–98. https://doi.org/10.1016/S2213-2600(19)30198-5.
Berry RB, Gamaldo CE, Harding SM, et al. The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications, Version 22. J Clin Sleep Med. 2015;11(403):1253–4. https://doi.org/10.5664/jcsm.5176.
Biard K, De Koninck J, Douglass AB. Creation of a shortened version of the sleep disorders questionnaire (SDQ). PLoS One. 2024;19(2):e0288216. https://doi.org/10.1371/journal.pone.0288216.
Cano-Pumarega I, Barbé F, Esteban A, Martínez-Alonso M, Egea C, Durán-Cantolla J. Sleep apnea and hypertension, are there sex differences? The Vitoria sleep cohort. Chest. 2017;152(4):742–50. https://doi.org/10.1016/j.chest.2017.03.008.
Casale M, Rinaldi V, Bressi F, et al. A suitable test for identifying high risk adult patients of moderate-severe obstructive sleep apnea syndrome. European Review for Medical and Pharmacological Sciences. 2008;12(4):275–80.
Chen R, Zhang Y, Ye W, et al. Improving OSA screening efficiency with subjective questionnaires: integrating STOP-Bang, ESS, and Berlin. Front Med. 2025;12:1581904. https://doi.org/10.3389/fmed.2025.1581904.
Chung F, Yegneswaran B, Liao P, et al. STOP questionnaire: a tool to screen patients for obstructive sleep apnea. Anesthesiology. 2008;108(5):812–21. https://doi.org/10.1097/ALN.0b013e31816d83e4.
Cleveland WS, Devlin SJ. Locally weighted regression: an approach to regression analysis by local fitting. J Amer Statistical Assoc. 1988;83(403):596–610.
Dodds S, Williams LJ, Roguski A, et al. Mortality and morbidity in obstructive sleep apnoea-hypopnoea syndrome: results from a 30-year prospective cohort study. ERJ Open Res. 2020. https://doi.org/10.1183/23120541.00057-2020.
Douglass AB, Bornstein R, Nino-Murcia G, et al. The sleep disorders questionnaire. I: creation and multivariate structure of SDQ. Sleep. 1994;17(2):160–7. https://doi.org/10.1093/sleep/17.2.160.
Douglass AB, Kaluzienski M. A simple manual neck examination predicts the apnea-hypopnea index obtained from polysomnography. Glob Transl Med. 2024. https://doi.org/10.36922/gtm.4548.
Duarte RL, Magalhaes-da-Silveira FJ, Oliveira ESTS, Silva JA, Mello FC, Gozal D. Obstructive sleep apnea screening with a 4-item instrument, named GOAL questionnaire: development, validation and comparative study with No-Apnea, STOP-Bang, and NoSAS. Nat Sci Sleep. 2020;12:57–67. https://doi.org/10.2147/NSS.S238255.
Duarte RLM, Magalhaes-da-Silveira FJ, Gozal D. Screening for obstructive sleep apnea: comparing the American Academy of Sleep Medicine proposed criteria with the STOP-Bang, NoSAS, and GOAL instruments. J Clin Sleep Med. 2023;19(7):1239–46. https://doi.org/10.5664/jcsm.10546.
Fabricant PD. Characteristics of a Diagnostic Test: Sensitivity, Specificity, Positive Predictive Value, and Negative Predictive Value. Chap. 5 in: Practical Clinical Research Design and Application: A Primer for Physicians, Surgeons, and Clinical Healthcare Professionals. Springer, Cham. 2024. pp. 31-36. https://doi.org/10.1007/978-3-031-58380-3_5.
Fedson AC, Pack AI, Gislason T. Frequently used sleep questionnaires in epidemiological and genetic research for obstructive sleep apnea: a review. Sleep Med Rev. 2012;16(6):529–37. https://doi.org/10.1016/j.smrv.2011.12.002.
Friedman M, Tanyeri H, La Rosa M, et al. Clinical predictors of obstructive sleep apnea. Laryngoscope. 1999;109(12):1901–7. https://doi.org/10.1097/00005537-199912000-00002.
Gamaldo C, Buenaver L, Chernyshev O, et al. Evaluation of clinical tools to screen and assess for obstructive sleep apnea. J Clin Sleep Med. 2018;14(7):1239–44. https://doi.org/10.5664/jcsm.7232.
Gauld C, Martin VP, Richaud A, et al. Systematic item content and overlap analysis of self-reported multiple sleep disorder screening questionnaires in adults. J Clin Med. 2023. https://doi.org/10.3390/jcm12030852.
Gauld C, Baillieul S, Martin VP, et al. Symptom content analysis of OSA questionnaires: time to identify and improve relevance of diversity of OSA symptoms? J Clin Sleep Med. 2024;20(7):1105–17. https://doi.org/10.5664/jcsm.11086.
Guilleminault C, Eldridge FL, Simmons FB, Dement WC. Sleep apnea in eight children. Pediatrics. 1976a;58(1):23–30.
Guilleminault C, Tilkian A, Dement WC. The sleep apnea syndromes. Annu Rev Med. 1976b;27:465–84. https://doi.org/10.1146/annurev.me.27.020176.002341.

Hersberger KE, Renggli VP, Nirkko AC, Mathis J, Schwegler K, Bloch KE. Screening for sleep disorders in community pharmacies - evaluation of a campaign in Switzerland. J Clin Pharm Ther. 2006;31(1):35–41. https://doi.org/10.1111/j.1365-2710.2006.00698.x.

Kim B, Lee EM, Chung YS, Kim WS, Lee SA. The utility of three screening questionnaires for obstructive sleep apnea in a sleep clinic setting. Yonsei Med J. 2015;56(3):684–90. https://doi.org/10.3349/ymj.2015.56.3.684.

Lee JH, Kim SJ, Jung HH. Nocturnal sleep related with metabolic markers in end-stage renal disease patients receiving hemodialysis. Psychiatry Investig. 2009;6(1):34–8. https://doi.org/10.4306/pi.2009.6.1.34.

Lomeli HA, Perez-Olmos I, Talero-Gutierrez C, et al. Sleep evaluation scales and questionnaires: a review. Actas Espanolas De Psiquiatria. 2007;36(1):50–9 https://actaspsiquiatria.es/index.php/actas/article/view/23.

Marti-Soler H, Hirotsu C, Marques-Vidal P, et al. The NoSAS score for screening of sleep-disordered breathing: a derivation and validation study. Lancet Respir Med. 2016;4(9):742–8. https://doi.org/10.1016/S2213-2600(16)30075-3.

Methipisit T, Mungthin M, Saengwanitch S, et al. The Development of Sleep Questionnaires Thai Version (ESS, SA-SDQ, and PSQI): Linguistic Validation, Reliability Analysis and Cut-Off Level to Determine Sleep Related Problems in Thai Population. J Med Assoc Thai. 2016;99(8):893–903 https://pubmed.ncbi.nlm.nih.gov/29947496/.

Miller JN, Berger AM. Screening and assessment for obstructive sleep apnea in primary care. Sleep Med Rev. 2016;29:41–51. https://doi.org/10.1016/j.smrv.2015.09.005.

Natsky AN, Vakulin A, Coetzer CLC, McEvoy RD, Adams RJ, Kaambwa B. Economic evaluation of diagnostic sleep studies for obstructive sleep apnoea: a systematic review protocol. Syst Rev. 2021;10(1):104. https://doi.org/10.1186/s13643-021-01651-3.

Netzer NC, Stoohs RA, Netzer CM, Clark K, Strohl KP. Using the Berlin questionnaire to identify patients at risk for the sleep apnea syndrome. Ann Intern Med. 1999;131(7):485–91. https://doi.org/10.7326/0003-4819-131-7-199910050-00002.

Perger E, Mattaliano P, Lombardi C. Menopause and sleep apnea. Maturitas. 2019;124:35–8. https://doi.org/10.1016/j.maturitas.2019.02.011.

Piperidou C, Karlovasitou A, Triantafyllou N, et al. Influence of sleep disturbance on quality of life of patients with epilepsy. Seizure. 2008;17(7):588–94. https://doi.org/10.1016/j.seizure.2008.02.005.

Schwegler K, Klaghofer R, Nirkko AC, Mathis J, Hersberger KE, Bloch KE. Sleep and wakefulness disturbances in Swiss pharmacy customers. Swiss Med Wkly. 2006;136(9–10):149–54. https://doi.org/10.4414/smw.2006.11265.

Seetho IW, Wilding JP. Screening for obstructive sleep apnoea in obesity and diabetes--potential for future approaches. Eur J Clin Invest. 2013;43(6):640–55. https://doi.org/10.1111/eci.12083.

Shi C, Wang Y, Luo J, Huang R, Xiao Y. Performance of four screening tools for identifying OSA among patients with insomnia. Nat Sci Sleep. 2025;17:379–90. https://doi.org/10.2147/NSS.S494804.

Takegami M, Hayashino Y, Chin K, et al. Simple four-variable screening tool for identification of patients with sleep-disordered breathing. Sleep. 2009;32(7):939–48 https://www.ncbi.nlm.nih.gov/pubmed/19639757.

Teodorescu M, Consens FB, Bria WF, et al. Predictors of habitual snoring and obstructive sleep apnea risk in patients with asthma. Chest. 2009;135(5):1125–32. https://doi.org/10.1378/chest.08-1273.

Valipour A, Lothaller H, Rauscher H, Zwick H, Burghuber CC, Lavie P. Gender-related differences in symptoms of patients with suspected breathing disorders in sleep: a clinical population study using the sleep disorders questionnaire. Sleep. 2007;30(3):312–9. https://doi.org/10.1093/sleep/30.3.312.

Valipour A, Lavie P, Lothaller H, Mikulic I, Burghuber OC. Sleep profile and symptoms of sleep disorders in patients with stable mild to moderate chronic obstructive pulmonary disease. Sleep Med. 2011;12(4):367–72. https://doi.org/10.1016/j.sleep.2010.08.017.

Vergara-Aguilar J-P, Silva-Soler MA, Vergara-Camargo JL, Dueñas-García FL. Sleep apnea syndrome in a Colombian population with epilepsy using the sleep apnea scale of the SDQ. Acta Neurol Colomb. 2023;39(4):1–10. https://doi.org/10.22379/anc.v39i4.1227.

Wang W, Yuan S, Le Grange JM, et al. Evaluating the performance of five scoring systems for prescreening obstructive sleep apnea-hypopnea syndrome. Sleep Breath. 2021;25(3):1685–92. https://doi.org/10.1007/s11325-020-02227-6.

Weatherwax KJ, Lin X, Marzec ML, Malow BA. Obstructive sleep apnea in epilepsy patients: the Sleep Apnea scale of the Sleep Disorders Questionnaire (SA-SDQ) is a useful screening instrument for obstructive sleep apnea in a disease-specific population. Sleep Med. 2003;4(6):517–21. https://doi.org/10.1016/j.sleep.2003.07.004.

Wong HB, Lim GH. Measures of diagnostic accuracy: sensitivity, specificity, PPV and NPV. Proc Singap Healthc. 2011;20(4):316–8. https://doi.org/10.1177/201010581102000411.

Xing Y, Zhang Z, Yin J, et al. Utilization of the GOAL questionnaire as a standardized screening tool for obstructive sleep apnea. Sci Rep. 2023;13(1):5489. https://doi.org/10.1038/s41598-023-31247-x.

Xu Z, Wu Y, Tai J, et al. Risk factors of obstructive sleep apnea syndrome in children. J Otolaryngol Head Neck Surg. 2020;49(1):11. https://doi.org/10.1186/s40463-020-0404-1.

Yu JL, Rosen I. Utility of the modified Mallampati grade and Friedman tongue position in the assessment of obstructive sleep apnea. J Clin Sleep Med. 2020;16(2):303–8. https://doi.org/10.5664/jcsm.8188.

Publisher's Note