Computerized Adaptive Testing for Sleep Disorders: Development of An Item Bank and Validation in A Simulated Study

Menghua She
Jiangxi Normal University https://orcid.org/0000-0002-8093-7762
Yaling Li
Jiangxi Normal University
Dongbo Tu
Jiangxi Normal University
Yan Cai (cy1979123@aliyun.com)
https://orcid.org/0000-0002-9406-1220

Research article

Keywords: sleep disorders; computerized adaptive testing; item response theory; IRT model; screening

Posted Date: April 22nd, 2020

DOI: https://doi.org/10.21203/rs.3.rs-18576/v1

Version of Record: A version of this preprint was published at European Journal of Health Psychology on July 1st, 2021. See the published version at https://doi.org/10.1027/2512-8442/a000076.

Abstract

Background: As more and more people suffer from sleep disorders, developing an efficient, cheap and accurate assessment tool for screening sleep disorders is becoming more urgent. This study developed a computerized adaptive testing for sleep disorders (CAT-SD).

Methods: A large sample of 1,304 participants was recruited to construct the item pool of CAT-SD and to investigate the psychometric characteristics of CAT-SD. More specifically, firstly the analyses of unidimensionality, model fit, item fit, item discrimination parameter and differential item functioning (DIF) were conducted to construct a final item pool which meets the requirements of item response theory (IRT) measurement. In addition, a simulated CAT study with real response data of participants was performed to investigate the psychometric characteristics of CAT-SD, including reliability, validity and predictive utility (sensitivity and specificity).

Results: The final unidimensional item bank of the CAT-SD not only had good item fit, high discrimination and no DIF; Moreover, it had acceptable reliability, validity and predictive utility.

Conclusions: The CAT-SD could be used as an effective and accurate assessment tool for measuring individuals' severity of the sleep disorders and offers a bran-new perspective for screening of sleep disorders with psychological scales.

Background

With the rapid development of the society, people's sleep problems are becoming more and more serious. According to the statistics of the World Health Organization (WHO), about 27 percent of the world's population have sleep disorders, and more than 300 million Chinese have sleep disorders (Wu, 2019). The rate of insomnia is 32–50% in the United States, 10–14% in Great British, 20% in Japan, 30% in France, and 38.2% in China (Zhang, 2015). In February 2018, the Philips company conducted sleep surveys of more than 15,000 people in 13 countries (United States, Britain, Germany, Poland, France, India, China, Australia, Colombia, Argentina, Mexico, Brazil and Japan) and found that a) most adults (67%) believe that sleep has a significant impact on health and well-being.; b) 61 percent of adults worldwide have some sort of medical problem that affects sleep (the Press Conference of 2018 World Sleep Day and the Launching Ceremony of the National Large-Scale Free Medical Consultation Activity, 2018). It can be concluded from the above surveys that sleep have become a very critical factor troubling people all over the world.

CCMD-3 defines non-organic sleep disorders as non-organic sleep and arousal disorders caused by a variety of psychosocial factors, including insomnia, narcolepsy, and certain paroxysmal sleep abnormalities (nightmare, restless legs, sleep walk et al.). Individuals with sleep disorders typically present with sleep-wake complaints of dissatisfaction regarding the quality, timing, and amount of sleep. Resulting daytime distress and impairment are core features shared by all of these sleep-wake disorders.

Sleep disorders have very negative effects on people's physical and mental health. Firstly, sleep disorders are often accompanied by depression, anxiety and cognitive changes, and studies have found that people who are chronically sleep-deprived are more likely to suffer from depression, anxiety and suicide than the general population (Jackson & Turkington, 2005; Joshua, 2014). Secondly, sleep disorders not only affect the treatment and rehabilitation of the primary disease, but also aggravate or induce other physical diseases, and bring greater

pain to patients. Thirdly, studies have proved that adequate sleep has a protective effect on the immune function of the body (e.g., Lange et al., 2010; Zai, 2017). After the occurrence of sleep disorders, the body's immune system is in an unbalanced state, which can lead to the reduction of immune cells and weakened immune function through a variety of ways. Finally, long-term sleep deprivation or lack will lead to the decline of people's attention, cognitive ability, judgment and memory, and even suffer from anxiety, depression, etc. At the same time, their daytime function is obviously affected, resulting in decreased work efficiency and living quality.

At present, the assessment methods of sleep disorders mainly include sleep diary, sleep behavior model, polysomnography (PSG) and sleep scales. Sleep diary is mainly used to record the subjects' sleep patterns during a period of time through continuous tracking, and then do a comprehensive analysis of their sleep status. Sleep behavior model is comprehensively evaluated by understanding and verification of the sleep-wake cycle, types and severity of sleep disorders of the subjects in detail. PSG integrates electroencephalograph (EEG), electrocardiogram (ECG), electrooculogram (EOG), electromyography (EMG) and other physiological detectors, it can collect various physiological changes during sleep. Sleep scales mainly require subjects to respond to the items in the scales, and then analyze the sleep status of the subjects according to their responses. For example, the Pittsburgh Sleep Quality Index (PSQI; Buysse et al., 1989) and the Insomnia Severity Index (ISI; Morin, 1993) are widely-used sleep scales.

Each of the above four sleep assessment methods has its own pros and cons. This article mainly introduces the shortcomings of sleep scales. Most of the existing sleep scales were developed within the framework of the classical test theory (CTT). One of the most prominent problems of CTT is that a large number of items are needed to cover a wide range of the construct with a high measurement precision.

In the clinical field, there is a high demand for mental health assessments which have both short duration and good quality (e.g., Gardner et al., 2004; Cella et al., 2007; Smits et al., 2007). Computerized Adaptive Testing (CAT) which involves the administration of items via the computer offers substantial promise for this. For CAT, each item is dynamically selected from item bank and is optimal for the respondents in question (Smits et al. 2011). CAT relies on modern test theory which is also known as Item Response Theory (IRT). The main content of IRT research is the relationship between the responses of subjects to items and the latent traits of subjects measured in the test. IRT models have item parameters which quantify the relationship between the latent trait and the item score (Smits et al. 2011). In recent years, a majority of researchers have used IRT to improving existing scales. For examples, O' Connor et al. (2014) used IRT to analyze the Subjective Happiness Scale (SHS), Cho et al. (2015) applied IRT to analyze emotional intelligence scale, and Wang (2018) developed an adaptive testing with a hierarchical item response theory (H-IRT) model. Generally speaking, CAT may be the most intriguing new perspective of IRT.

Compared with traditional Paper & Pencil (P&P) test, the greatest advantage of CAT is that it can greatly reduce the number of items without loss of measurement accuracy. In addition, CAT has many other advantages. For instance, the presentation of items is more standardized, not only can accurately control what the examinee can see and hear, but can control the length of time. However, CAT also has disadvantages, such as being a complex technique and requiring a substantial amount of human and financial resources to organize a CAT program. However, study had shown that the advantages of CAT far outweigh its disadvantages (Meijer & Nering, 1999).

According to literature review, CAT has been widely used in psychological and clinical fields. For examples, Fliege et al. (2005) developed a CAT for depression; Abberger et al. (2013) developed a CAT to assess anxiety in

cardiovascular rehabilitation patients; Gibbons et al. (2017) used a CAT of the quality of life to adjust the cross-cultural differences of the participants. However, few CAT studies for sleep disorders have been found, which is unfavorable to the measurement and assessment of sleep disorders. Considering the seriously negative effects of sleep disorders on people, so it is urgent to develop an efficient and accurate assessment tool for sleep disorders. Moreover, the technology, algorithm and implementation of CAT for sleep disorders deserve further discussion. To address the above issues, this study attempts to propose a CAT using real data to measure sleep disorders (CAT-SD).

The current study is expected to contribute to the theory and practice to the measurement and assessment of sleep disorders. Specifically, in theory, a) the combination of CAT and sleep disorders broadens the application area of CAT; b) although some sleep scales have short items, these scales may not adequately refer to the domains related to sleep disorders. However, multiple sleep scales are combined into a large item bank which is relatively comprehensive in CAT; c) CAT mitigates measurement error while maximizing efficiency since only the items pertinent to accurately measuring trait level are administered (Kirisci et al., 2012). In practice, a) cost is minimal because the responses are scored automatically and immediately after the subject complete the questionnaire; b) privacy is also ensured because there is no record of the subjects' responses on paper and access to the information is protected by password (Kirisci et al., 2012); c) the sleep scales are very important as an auxiliary tool for doctors to diagnose patients' sleep status, and the patients can complete it at home.

The next is to describe the development of CAT-SD and the evaluation of its test properties in simulation studies.

Methods

Participants

The total college student sample came from ten universities in five Chinese cities (Beijing, Shanghai, Jingzhou, Jingdezhen and Nanchang) that consisted of 1,304 participants who agreed to take part in this study after being informed that their personal information would be kept secret, all participants were voluntary and anonymous. Participants included both healthy individuals (81.18%) and patients (18.82%). These patients reported that they had been diagnosed with sleep disorders by professional doctors and were screened with severe sleep disorders by several sleep scales. The sample comprised 585 males (44.86%) and 719 females (55.14%), of whom 752 (57.67%) were from rural areas and 552 (42.33%) were from cities, and the mean age was 19.15 years old (SD = 1.52, range from 15 to 25), 87.40% of participants aged from 18 to 22.

Measures

The first step is to conduct a Delphi process. We started with 133 available items originating from 8 self-rating sleep scales that are widely-used in routine diagnostic examinations: Pittsburgh Sleep Quality Index (PSQI; Buysse et al., 1989), Insomnia Severity Index (ISI; Morin, 1993), Epworth Sleepiness Scale (ESS; Johns, 1991), SLEEP-50 questionnaire (SLEEP-50; Spoormaker et al., 2005), Self-Rating Scale of Sleep (SRSS; Li, 2012), Chinese Sleep Disorder Scale (CSDS; Zhang, 2014), Medical Outcomes Study Sleep Scale (MOS-SS; Hays et al., 2005), and Quality of Life in Neurological Disorders-Sleep Disturbance Scale (Neuro-QOL-SD; Perez et al., 2007). Athens Insomnia Scale (AIS; Soldatos et al., 2000) which is widely-used in diagnosing sleep disorders is selected as criterion scale to evaluate the CAT-SD. These questionnaires are administered to participants via computer.

Statistical Analysis

Statistical analyses are mainly composed of two parts: construction of item bank for CAT-SD, and the CAT-SD simulation study.

Construction of item bank for CAT-SD

This part attempts to construct an item bank that meets the requirements of IRT measurement. The statistical analyses based on IRT were sequentially carried out, including unidimensionality, model fit, item fit, discrimination parameters and differential item function (DIF).

Unidimensionality. Unidimensionality is a crucial assumption in IRT, which means a test only measures one main latent trait, and no other factors will affect the characteristics of the examinee's response to items. Many IRT models assumed unidimensional, such as the two-parameter Logistic model (2PL), three-parameter Logistic model (3PL), the Graded Response Model (GRM; Samejima, 1968) and the Generalized Partial Credit Model (GPCM; Muraki, 1992). To confirm acceptable unidimensionality of the dataset, EFA was conducted. If the ratio of the first eigenvalue to the second eigenvalue is greater than 4 (Reeve et al., 2007) along with the first eigenvalue explains more than 20% of the total variance (Reckase, 1979), which can be considered that the test conforms to the unidimensional hypothesis.

Model fit. In IRT, selecting an optimal IRT model for statistical analyses is the premise to ensure the accuracy of statistical analyses. In current study, four polytomously-scored IRT models (i.e., GRM, PCM and GPCM) were simultaneously applied to fit the items of CAT-SD, and the optimal model was selected based on the test-level model-fit indices include − 2Log-Likelihood (-2LL; Spiegelhalter et al., 1998), Akaike's information criterion (AIC; Akaike, 1974) and the Bayesian information criterion (BIC; Schwarz, 1978). Smaller values of these indices indicate better model-fit.

Discrimination parameters. Item discrimination parameters (*a*) shows the extent to which individuals with similar scores can be differentiated via an item. An item with high discrimination parameter suggests that this item is helpful to obtain more precise estimation of examinee's latent traits. Item discrimination parameters were estimated via the optimal model and then selected items with discrimination parameters more than 0.5 (Chang & Ying, 1996).

Item fit. In order to calculate the item fit, the $S\text{-}X^2$ statistic (Kang & Chen, 2008) that quantifies and compares the differences between observed frequencies and expected frequencies under the IRT model was suggested. Items with p values of $S\text{-}X^2$ less than 0.01 were deemed to have poor item-fit (Flens et al., 2017) and were removed.

Differential item function (DIF). Another assumption of IRT models is that the item has same item parameters in different sample. If parameter values differ between groups, a test is said to suffer from Differential Item Function (DIF; Embretson & Reise, 2000, Chap. 10). The consequence of DIF is that respondents from different groups, who actually have an identical score on the latent trait, have a different probability of endorsing an item (Smits et al., 2011). Change in McFadden's pseudo R^2 was used to evaluate effect size, and the null hypothesis of no DIF was rejected when the value of R^2 change was more than 0.02 (Bjorner, 2003). DIF analysis was carried out with respect to gender (male, female) and region (rural, city) groups.

The IRT analyses of unidimensionality, model fit, item fit, discrimination parameters and DIF were sequentially performed until the remaining items of item bank fully satisfied the above rules. Consequently, the remaining

items constructed the final item bank of CAT-SD, and then the item parameters of the final item bank and person parameters were estimated against by using the optimal model.

CAT-SD Simulation Study

In this part, all the real participants' response data were used for a CAT-SD simulation study that was to investigate the characteristic, criterion-related validity and predictive utility (sensitivity and specificity) of CAT-SD.

Initial item selection. In the CAT simulation, item selection is dependent on the participants' responses to a given item. However, the computer knows nothing about prior information of participants at the beginning. the random selection method was applied to initial item selection in this study.

Item selected method. Once the CAT has an estimate of participants' latent trait, it needs according to the estimated latent trait to choose the most appropriate item for them. The Maximum Fisher Information (MFI) (Baker, 1992) method is a commonly used method for item selection by selecting the item with maximum information at that estimated theta point. This so-called statistical information is a function of the item parameters and is related to the measurement error of the estimated latent variable. The higher the information of the item, the more it reduces the measurement error associated with that estimate (Smits et al., 2011). The Fisher Information was defined as,

$$I_j(\hat{\theta}) = \sum_{k=1}^{K} \frac{\left[P_k^{'}(\hat{\theta})\right]^{\wedge 2}}{P_k(\hat{\theta})}$$

where the $I_j(\hat{\theta})$ is the item information function of item *j* given the θ, the θ is the estimated latent trait level, $P_k(\hat{\theta})$is the probability of getting score *k* given the θ, and $P_k^{'}(\hat{\theta})$ is the first derivative of $P_k(\hat{\theta})$ to θ. A new item with the highest $I_j(\hat{\theta})$ at that estimated theta point was selected.

Scoring algorithm. After the participants responded to an item, their sleep disorders theta was updated with the expected a posteriori method (EAP; Bock & Mislevy, 1982). EAP is a kind of Bayesian estimation based on the participants' responses to the selected item. The EAP was defined as,

$$\hat{\theta}_i = \frac{\sum_{k=1}^{q} \theta_k L_i(\theta_k) A(\theta_k)}{\sum_{k=1}^{q} L_i(\theta_k) A(\theta_k)}$$

where the θ_k refers to be one of the θ quadrature points, $L_i(\theta_k)$is the likelihood function of participant *i* with a specific response, given an ability value θ_k $A(\theta_k)$ is the weight of the quadrature points, and $\sum_{k=1}^{q} A(\theta_k)=1$.

Stopping rule. The CAT algorithm alternately selects items and updates the estimate of the participant's latent trait until the item bank is empty, unless termination criterions are set. There are generally two approaches to terminate the test, one is fixed length, the other is variable length. In this study, the latter rule was applied, that is, the test was terminated when the standard error (*SE*) of theta reached the pre-set value of *SE(θ)*. *SE* for a trait level (Magis & Raiche, 2012) can be defined as,

$$SE(\theta) = \frac{1}{\sqrt{\sum_{j=1}^{n} I_j(\theta)}}$$

where the *n* denoted the total number of administrated items. Different stopping rules have different estimation accuracy for the test. Generally speaking, the larger the *SE*(*θ*), the lower the accuracy of the test estimate, and vice versa. Several cut-off values of *SE(θ)* were used in the CAT-SD simulation: the whole final item bank (None), *SE(θ)* ≤ 0.3, *SE(θ)* ≤ 0.4, *SE(θ)* ≤ 0.5, and *SE(θ)* ≤ 0.6, respectively (Tan, et al., 2018).

Characteristic of CAT-SD. In order to explore the characteristics of the CAT-SD, several statistics were calculated: the mean and standard deviation (*SD*) of administrated items, the mean *SE* of theta estimates, the Pearson's correlation between the estimated theta under different stopping rules and theta estimations using the whole item bank, and the marginal reliability that was the mean reliability for all levels of theta (Smits et al., 2011). The corresponding reliability of each examinee can be derived via the following formula (Samejima, 1994) when the mean and *SD* of theta were 0 and 1, respectively,

$$r_{xx}(\theta_i) = 1 - \frac{1}{I(\theta_i)}$$

where the $I(\theta_i)$is the test information for the participant , which can be inferred based on the administered item parameters and his/her respond, the $r_{xx}(\theta_i)$ is the corresponding reliability in IRT for the participant , and the marginal reliability is the average of the corresponding reliability of each participant. Additionally, the figure of the number of selected items and test information as functions of estimated theta under different stopping rules was plotted. The test information suggests the measurement precision of CAT-SD, and the lower the value of it, the larger the error of the theta estimation.

Criterion-related validity and Predictive utility (sensitivity and specificity) of CAT-SD. In order to further investigate the criterion-related validity and predictive

utility (sensitivity and specificity) of CAT-SD, the AIS that is widely-used and well-validated in measuring sleep disorders was selected as criterion scale. The Pearson's correlations between the estimated theta in the CAT-SD and the standard scores of the AIS under different stopping rules were calculated. Predictive utility (sensitivity and specificity) of CAT-SD was examined calculating Receiver Operating Characteristics (ROC). The area under the curve (AUC) can be seen as the probability that a randomly selected unhealthy individual scores higher than a randomly selected healthy individual on the sleep scales. and its value ranged from 0.5 to 1. The predictive utility of the estimated theta for diagnosing sleep disorders is similar to random guessing when AUC = 0.5, while it is optimal when AUC = 1. Swets and colleagues (1988) suggested to heuristically interpret AUC-values as small (0.5 < AUC $\leqslant$ 0.7), moderate (0.7 < AUC $\leqslant$ 0.9), or high (0.9 < AUC $\leqslant$ 1). Sensitivity refers to the probability that a patient is accurately diagnosed with a disease, and specificity refers to the probability that healthy individual is diagnosed with no illness, the larger the value of these two indicators, the better the effect of the diagnosis (Tan, et al., 2018). Determination of the cut-off scores was calculated by maximizing the Youden-Index (YI = sensitivity + specificity − 1) (Schisterman et al., 2005). The AIS regarded as the classified variable (according to the scoring standard of this scale, when the total score of the participants was greater than 6, they were diagnosed as insomnia and were rated as 1, while others were rated as 0) and the estimated theta in CAT-SD was used as a continuous variable for sleep disorders to plot the ROC curve under different stopping rules.

Software

The EFA and ROC curve were carried out via the software SPSS 23.0. All other analyses were performed in the free statistical package R (Version 3.4.1; Coreteam, 2015). Specifically, the analyses of IRT model selection, item fit and discrimination parameters were conducted via the 'mirt' package (Version 1.24; Chalmers, 2012); DIF tests via the 'lordif' package (Version 0.3-3; Choi, 2015); the 'catR' package (Version, Magis & Barrada, 2017) was applied to conduct CAT algorithm.

Results

Construction of item bank for CAT-SD

Unidimensionality

Results of EFA for 133 items in the initial item bank for sleep disorders indicated that the ratio of the first eigenvalue (λ_1 = 22.003) to the second eigenvalue (λ_2 = 4.552) was 4.834, and the first eigenvalue explains accounted for 24.448% of the total variance, which was more than 20% of the total variance. Thereby, these results revealed the unidimensionality of CAT-SD.

Model fit

Model fit statistical indices of the GRM, the GPCM and the PCM were documented in Table 1. It can be seen from the table that the values of -2LL, AIC and BIC of GRM model were all smaller than those of other IRT models which suggested that the GRM fitted the data best. Therefore, the GRM was regarded as optimal model to perform the further analyses.

Table 1
Test-level model-fit for four polytomously-scored IRT models.

Model	*-2LL*	*AIC*	*BIC*
GRM	331495.4	332651.4	335641.6
GPCM	333172.8	334328.9	337319.0
PCM	353714.6	355071.8	355497.0

Note: -2LL= -2Log-Likelihood; AIC = Akaike' information criterion; BIC = Bayesian information criterion; GRM = Graded Response Model; GPCM = Generalized Partial Credit Model; PCM = Partial Credit Model.

Discrimination parameters, Item fit, and DIF

The discrimination parameter of 13 items in the initial item bank were less than 0.5, so they would be removed from the item bank. Of the remaining 120 items, the $S\text{-}X^2$ values of 14 items were less than 0.01, therefore, the 14 items were removed due to poor item-fit. Regarding DIF of the remaining 106 items, there was no DIF in the region group, while there were 12 items that the values of R^2 change were more than 0.02 in the gender group, thus, the 12 items with DIF were eliminated.

Consequently, the final item bank of CAT-SD comprised 94 items after 39 items were eliminated for the above psychometric criterions. After that, unidimensionality, model fit, item fit, discrimination analysis and DIF test were

conducted again for the remaining 94 items, and it was concluded that all the items met the requirements of IRT measurement. The statistical indices of items in the final item bank of CAT-SD were partly presented in Table 2 and the statistics of the whole item bank were provided in the Supplementary material. For the final item bank, the average IRT discrimination parameter (*a*) was 1.31 (*SD* = 0.40), which suggested the final item bank had high quality, and the location parameter (*b*) ranged from − 4.66 to 4.88, which implied the location parameter basically covered a large range of traits.

Table 2
Fundamental information of part items in the final item bank for CAT-SD (N = 94).

Item	*Abbreviated item content*	*Item parameters*							*Item-fit estimates DIF*			
		a	Grade	b1	b2	b3	b4	b5	S-X2	df	p value	R^2 change
N2	Wake up early	0.95	4	-0.05	1.43	3.24	-	-	251.57	252	0.496	0.0005
N3	Go to the bathroom	0.63	4	0.23	2.55	4.28	-	-	246.77	259	0.697	0.0004
N4	Not breathing properly	1.44	4	0.90	1.94	3.36	-	-	172.71	161	0.250	0.0082
N5	Coughing or snoring loudly	0.76	4	0.99	2.53	4.65	-	-	261.75	228	0.062	0.0078
N6	Feel cold	0.71	4	-0.20	1.44	3.43	-	-	174.06	184	0.689	0.0022
N7	Feel hot	0.87	4	0.36	1.94	4.28	-	-	250.53	246	0.408	0.0018
N8	Having nightmares	0.90	4	-0.06	1.86	4.17	-	-	274.59	243	0.080	0.0014
N9	Pain and discomfort	1.38	4	0.85	2.00	3.36	-	-	158.52	161	0.541	0.0099
N11	Other things that interfere with sleep	1.12	4	-0.67	0.82	2.49	-	-	258.98	253	0.385	0.0069
N12	Sleep quality	1.31	4	-1.44	1.11	3.04	-	-	174.36	199	0.896	0.0106
N15	Have insufficient energy to do things	1.15	4	-1.44	1.15	2.55	-	-	224.94	231	0.600	0.0126

CAT-SD Simulation Study

Characteristic of CAT-SD

Table 3 shows several characteristics of CAT-SD under different stopping rules. The first row shows the characteristics of CAT-SD when no stopping rule was applied, that is, all items in the final item bank were

administered. The second and third columns show the mean number of items administered and the associated *SD*, respectively. Obviously, the higher the level of measurement precision, the more the mean number of used items. The fourth column shows the average *SE* of the estimated theta of each examinee for each stopping rule. The fifth and sixth columns show the marginal reliabilities and the Pearson's correlation coefficients, the marginal reliabilities with an average of 0.84 (range from 0.70 to 0.97), and the Pearson's correlation coefficients range from 0.86 to 1 under different stopping rules. Evidently, when the stopping rules were set below 0.4, the marginal reliabilities and Pearson's correlation coefficients were very high.

Table 3
Characteristics of the CAT-SD under different stopping rules.

Stopping rule	Number of items used		*Mean SE(θ)*	*Marginal reliability*	*r*
	Mean	SD			
None	94	0	0.16	0.97	1.00**
$SE(\theta) \leq 0.6$	3.89	2.40	0.54	0.70	0.86**
$SE(\theta) \leq 0.5$	5.55	4.72	0.46	0.79	0.89**
$SE(\theta) \leq 0.4$	8.47	6.76	0.38	0.85	0.91**
$SE(\theta) \leq 0.3$	15.19	9.09	0.30	0.91	0.95**

Note: ** shows the discrepancy on 0.01 level notable. None = the whole item bank was administered; r = the Pearson's correlations between the estimated theta in the CAT-SD and the estimated theta via the whole item bank.

Figure 1 depicts the standard error of estimated theta of the CAT-SD under stopping rules $SE(\theta) \leq 0.4$ and $SE(\theta) \leq 0.3$. Subjects who have moderate and high CAT-SD score have a smaller standard error. The results were consistent with the fact that screening was more effective for people with moderate or severe sleep disorders than for people with mild sleep disorders. Figure 2 shows the number of items administered along with test information as functions of the estimated theta in the CAT-SD with stopping rules $SE(\theta) \leq 0.4$ and $SE(\theta) \leq 0.3$. Particularly, a large number of items had to be administered for subjects with lower theta and the test information was low. However, fewer items were administered for subjects with middle or high theta and the test information was high. For example, under the stopping rule $SE(\theta) \leq 0.3$,a) the test information was less than 8 for those whose theta ranged from − 3.8 to − 2 even if the entire item bank was administered to them; while b) the test information was over 12 for those whose theta ranged from 0 to 3 with about 10 administered items to them. Figure 3 illustrates the density distributions of the sleep disorder scores obtained from traditional test (the whole item bank) and CAT-SD. As we can see, the two distributions are almost identical with different stopping rules, the Pearson's correlations between two kinds of test are 0.91 and 0.95 under stopping rules $SE(\theta) \leq 0.4$ and $SE(\theta) \leq 0.3$ (see Table 4). Figure 4 displays the marginal reliabilities of CAT-SD for each participant with different stopping rules. Under $SE(\theta) \leq 0.4$ and $SE(\theta) \leq 0.3$, the marginal reliabilities were above the average of it ($r = 0.84$), which indicated that CAT-SD had a high reliability for most participants. Furthermore, the marginal reliabilities for participants with estimated theta more than − 2.5 under $SE(\theta) \leq 0.3$ was maximal, while the marginal reliabilities under $SE(\theta) \leq 0.3$ and $SE(\theta) \leq 0.4$ were equal when estimated theta was less than − 2.5, and the marginal reliabilities under $SE(\theta) \leq 0.3$, $SE(\theta) \leq 0.4$ and $SE(\theta) \leq 0.5$ were equal when estimated theta was less than − 3. Individuals always had the minimum marginal reliabilities with stopping rule $SE(\theta) \leq 0.6$, regardless of theta estimation.

Criterion-related validity and Predictive utility (sensitivity and specificity) of CAT-SD

The results show that the correlations (range from 0.684 to 0.777, $p < 0.001$) between CAT-SD estimated theta and AIS was significant under different stopping rules, which indicated the CAT-SD had acceptable criterion-related

validity. The ROC analysis for CAT-SD is presented in Table 4 and Fig. 5. It can be clearly seen that the ROC curves under the five stopping rules are very close. AUC = 90.2% (95% confidence interval= [88.5%, 91.9%]), sensitivity = 79.5% and specificity = 86.5%, when no stopping rule was applied. Then the value of AUC dropped to 85.7%, (95% confidence interval= [83.7%, 87.8%]), sensitivity = 79.9% and specificity = 76.1% under stopping rule of SE(θ) ≤ 0.6. Even so, the values of AUC were also higher than critical value 0.7 that is universally regarded as the lower bound for moderate predictive utility under all stopping rules. In this study, the minimum probability that patients were accurately screened with sleep disorders and that normal individuals were accurately screened with no sleep disorders were 0.781 and 0.751, which were higher than the random level (0.5).

Table 4
The predictive utility (sensitivity and specificity) of the CAT-SD under different stopping rules.

Stopping rule	***AIS***				
	AUC (95% CI)	**Cut-off**	**Se**	**Sp**	**YI**
None	0.902(0.885–0.919)	0.314	0.795	0.865	0.660
SE(θ) ≤ 0.6	0.857(0.837–0.878)	0.089	0.799	0.761	0.560
SE(θ) ≤ 0.5	0.862(0.842–0.882)	0.093	0.815	0.751	0.566
SE(θ) ≤ 0.4	0.862(0.842–0.882)	0.209	0.793	0.782	0.575
SE(θ) ≤ 0.3	0.878(0.859–0.897)	0.264	0.781	0.813	0.594
Note: 95% CI = 95% confidence interval; None = the whole item bank was administered; AUC = Area Under Curve; Se = Sensitivity; Sp = Specificity; YI = Youden-Index.					
[insert Fig. 5.]					

Discussion

This study focused on the development of CAT-SD, which provided optimal items for individuals based on the severity of their sleep disorders to effective assessment sleep disorders and significantly reduce the test burden without loss of measurement accuracy.

The whole study was divided into two parts: construction of the CAT-SD item bank and CAT-SD simulation study. In order to construct a high-quality item bank for CAT-SD, items were carefully selected from eight universally-used sleep scales. After the unidimensionality, model fit, item fit, discrimination analysis and DIF test were carried out, a high-quality item bank was constructed. Results display that the final unidimensional item bank of CAT-SD contained 94 items which had good item-fit, high discrimination and no DIF. In CAT-SD simulation study, the real participants' response data was used to investigate the psychometric characteristics of CAT-SD, including reliability, validity and predictive utility (sensitivity and specificity). Simulated CAT-SD under different stopping rules (required standard errors in decreasing steps of 0.1) were performed and results revealed that, a) individuals with moderate or severe sleep disorders can be accurately screened by administering only a few items. Small differences are more easily detected for participants with high scores than those with low scores of sleep disorders in the context of participants with a similar degree of sleep disorders. This result is similar to previous studies (e.g. Smits et al., 2011, Reise & Waller, 2009); b) high correlation were observed between the traditional test and the CAT-SD. But participants only need to complete 15.19 and 8.47 items under the stopping rules *SE(θ)* ≤ 0.3

and $SE(\theta) \leq 0.4$ (see Table 4) in CAT-SD. CAT offers main advantages to the traditional test is that only the optimal items are administrated to each participant, minimizing test burden without sacrificing measurement precision; c) CAT-SD had an acceptable marginal reliability with an average of 0.84. Meanwhile, it also had an acceptable and reasonable criterion-related validity with the AIS, the Pearson's correlation coefficients under different stopping rules were all greater than 0.6 that is widely-used as the lower bound for moderate correlation; d) from the ROC curve analysis, the AUC values (AUC = 0.857 ~ 0.902) did not change much and were higher than the value (0.7) of the lower bound for a moderate predictive utility under different stopping rules, therefore, the CAT-SD had a good screening performance for sleep disorders. The sensitivity (0.781 ~ 0.815) and specificity (0.751 ~ 0.865) of the CAT-SD were both acceptable; e) the simulation study in this study indicated that the stopping rules $SE(\theta) \leq 0.3$ and $SE(\theta) \leq 0.4$ seem to be optimal. Because the two stopping rules are higher than the other stopping rules in terms of reliability and validity, although more items were used, the number of administrated items was within the acceptable range.

Although appreciating the promising results for the proposed CAT-SD, there were still some limitations. Firstly, when the criterion-related validity and the predictive utility (sensitivity and specificity) of CAT-SD were performed, only one scale was selected as the criterion scale and the participants were same. Future studies should employ more criterion scales to stabilize and cross-validate the validity of CAT-SD and to ensure that the subjects involved in experiment and verification process are different. Secondly, given that the CAT item bank requires a sufficient number of items with high quality and a wide range of location parameters (Howard, 1990), more high quality items should be supplemented to the item bank of CAT-SD in future. The size of the item bank that is generally considered appropriate should be 6 to 12 times the number of items in P&P test (Stocking et al., 1993). There were 96 items in the final item bank of CAT-SD, but if the item exposure rate, item elimination, item content distribution and other issues are taken into consideration, the existing item bank should be further expanded. Thirdly, the sample is not representative. Therefore, future researches need to be done on participants with sleep disorders. Finally, in current research, CAT-SD simulation study with real response data in traditional test was carried out, however, a real CAT-SD administration should be implemented in future researches to further explore the efficiency of CAT-SD. Different results may be produced by simulated and real CAT administration (Smits et al., 2011). In real situation, the participants' responses will be affected by many factors, such as the environment, mood, people and time, however, simulation study were usually performed under ideal conditions. Fortunately, research (Kocalevent et al., 2009) had found that the results of simulated CAT were consistent with actual CAT. Consequently, this study still has some practical significance.

Conclusion

The CAT-SD could be used as an effective and accurate assessment tool for measuring individuals' severity of the sleep disorders and offers a bran-new perspective for screening of sleep disorders with psychological scales. We need more studies to assess the performance of adaptive tests in both mental health specialty and other clinical settings.

List Of Abbreviations

AIC: Akaike's information criterion	MFI: Maximum Fisher Information
AIS: Athens Insomnia Scale	ML: Maximum Likelihood
AUC: Area Under the Curve	MOS-SS: Medical Outcomes Study Sleep Scale
BIC: Bayesian information criterion	Neuro-QOL-SD: Quality of Life in Neurological Disorders-Sleep Disturbance Scale
CAT-SD: Computerized Adaptive Testing for Sleep Disorders	NREM: Non-Rapid Eye Movement
CFA: Confirmatory Factor Analysis	NRM: Nominal Response Model
CTT: Classical Test Theory	PCM: Partial Credit Model
CSDS: Chinese Sleep Disorder Scale	P&P: Paper & Pencil
DIF: Differential Item Functioning	PSG: polysomnography
ECG: electrocardiogram	PSQI: Pittsburgh Sleep Quality Index
EEG: electroencephalograph	REM: Rapid Eye Movement
EFA: Exploratory Factor Analysis	ROC: Receiver Operating Characteristic
EMG: electromyography	RSM: Rating Scale Model
EOG: electrooculogram	SE: Standard Error
ESS: Epworth Sleepiness Scale	SHS: Subjective Happiness Scale
GPCM: Generalized Partial Credit Model	SLEEP-50: SLEEP-50 questionnaire
GRM: Graded Response Model	SRSS: Self-Rating Scale of Sleep
H-IRT: Hierarchical Item Response Theory	2PL: Two-Parameter Logistic model
IRT: Item Response Theory	3PL: Three-Parameter Logistic model
ISI: Insomnia Severity Index	-2LL: -2Log-Likelihood

Declaration

—Ethics approval and consent to participate

The study was carried out following the recommendations of psychometrics studies on mental health at the Research Center of Mental Health, Jiangxi Normal University and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. All participants and their parents or legal guardian provided verbal informed consent and this practice was approved by the ethics committee of the Research Center of Mental Health, Jiangxi Normal University.

—Consent for publication

Not applicable

—Availability of data and materials

AIC: Akaike's information criterion MFI: Maximum Fisher Information
The data and materials that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.
—Competing interests
All authors declare that they have no any conflict of interest related to this work.
—Funding
This work was supported by the National Natural Science Foundation of China [grant numbers 31960186, 31760288, 31660278]. The funders had no role in study design, data collection, data analysis, data interpretation, or writing of the manuscript.

Declarations

Ethics approval and consent to participate

The study was carried out following the recommendations of psychometrics studies on mental health at the Research Center of Mental Health, Jiangxi Normal University and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. All participants and their parents or legal guardian provided verbal informed consent and this practice was approved by the ethics committee of the Research Center of Mental Health, Jiangxi Normal University.

Consent for publication

Not applicable

Availability of data and materials

The data and materials that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Competing interests

All authors declare that they have no any conflict of interest related to this work.

Funding

This work was supported by the National Natural Science Foundation of China [grant numbers 31960186, 31760288, 31660278]. The funders had no role in study design, data collection, data analysis, data interpretation, or writing of the manuscript.

Authors' contributions

MHS conceived the study design, analyzed the data and wrote the first draft of the manuscript. YLL conducted data collection and literature search. DBT revised the draft. YC collected and assessed participants. All authors read and approved the manuscript.

Acknowledgements

The authors are grateful to all the participants for their cooperation in our study. We also sincerely thank the teachers of Jiangxi Normal University, the doctors of Jiangxi Mental Hospital and the Second Affiliated Hospital of Nanchang University.

References

Abberger, B., Haschke, A., Wirtz, M., Kroehne, U., Bengel, J., & Baumeister, H. Development and Evaluation of a Computer Adaptive Test to Assess Anxiety in Cardiovascular Rehabilitation Patients. *Archives of Physical Medicine and Rehabilitation*. 2013; 94(12):2433–2439. doi:10.1016/j.apmr.2013.07.009

Akaike, H. A new look at the statistical model identification. *Automatic Control IEEE Transactions on.* 1974; 19(6):716–723. doi:10.1109/tac.1974.1100705

Andrich, D. A rating formulation for ordered response categories. *Psychometrika*. 1978; 43(4):561–573. doi:10.1007/bf02293814

Baker, F. B. *Item Response Theory: Parameter Estimation Techniques*. New York, NY: Marcel Dekker; 1992.

Bjorner, J. B., Kosinski, M., Ware, J. E. Calibration of an Item Pool for Assessing the Burden of Headaches: An Application of Item Response Theory to the Headache Impact Test (HIT™). *Quality of Life Research*. 2003; 12(8):913-933. doi:10.2307/4038949

Darrell Bock, R. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*. 1972; 37(1):29–51. doi:10.1007/bf02291411

Bock, R. D., & Mislevy, R. J. Adaptive EAP Estimation of Ability in a Microcomputer Environment. *Applied Psychological Measurement*. 1982; 6(4):431–444. doi:10.1177/014662168200600405

Buysse, D. J., Reynolds, C. F., Monk, T. H., Berman, S. R., & Kupfer, D. J. The Pittsburgh sleep quality index: A new instrument for psychiatric practice and research. *Psychiatry Research*. 1989; 28(2):193–213. doi:10.1016/0165-1781(89)90047-4

Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B., Ader, D., Fries, J., Bruce, B., & Rose, M. The patient-reported outcomes measurement information system (PROMIS): progress of an NIH roadmap cooperative group during its first two years. *Medical Care*. 2007; 45:S3-S11. doi: 10.1097/01.mlr.0000258615.42478.55

Chalmers, R. P. mirt: a multidimensional item response theory package for the r environment. *Journal of Statistic Software*. 2012; 48:1-29. doi:10.18637/jss.v048.i06

Chang, H. H., & Ying, Z. A global information approach to computerized adaptive testing. *Applied Psychological Measurement*. 1996; 20(3):213-229. doi:10.1177/01466216960200030

Cho, S., Drasgow, F., & Cao, M. An investigation of emotional intelligence measures using item response theory. *Psychological Assessment*. 2015; 27(4):1241–1252. doi:10.1037/pas0000132

Choi, S. W. *Lordif: Logistic Ordinal Regression Differential Item Functioning Using IRT*; 2015.

Coreteam, R. R: a language and environment for statistical computing. *Computing*. 2015; 14:12-21. doi:10.1890/0012-9658(2002)083[3097:CFHIWS]2.0.CO;2

Embretson, S., & Reise, S. P. *Item Response Theory for Psychologists*. Lawrence Erlbaum, Mahwah, NJ; 2000.

Flens, G., Smits, N., Terwee, C. B., Dekker, J., Huijbrechts, I., & de Beurs, E. Development of a Computer Adaptive Test for Depression Based on the Dutch-Flemish Version of the PROMIS Item Bank. *Evaluation & the Health Professions*. 2017; 40(1):79–105. doi:10.1177/0163278716684168

Gardner, W., Shear, K., Kelleher, K. J., Pajer, K. A., Mammen, O., Buysse, D., & Frank, E. Computerized adaptive measurement of depression: A simulation study. *BMC Psychiatry*. 2004; 4(1). doi:10.1186/1471-244x-4-13

Gibbons, C. J., & Skevington, S. M. Adjusting for cross-cultural differences in computer-adaptive tests of quality of life. *Quality of Life Research*. 2017; 27(4):1027–1039. doi:10.1007/s11136-017-1738-7

Hays, R. D., Martin, S. A., Sesti, A. M., & Spritzer, K. L. Psychometric properties of the Medical Outcomes Study Sleep measure. *Sleep Medicine*. 2015; 6(1):41–44. doi:10.1016/j.sleep.2004.07.006

Howard, W. *Computerized adaptive testing: A primer*. Hillsdale, NJ: Lawrence Erlbaum Associates; 1990

Jackson, M. J. Depression and anxiety in epilepsy. *Journal of Neurology, Neurosurgery & Psychiatry*. 2005; 76(suppl_1): i45–i47. doi:10.1136/jnnp.2004.06046

Johns, M. W. A New Method for Measuring Daytime Sleepiness: The Epworth Sleepiness Scale. *Sleep*. 1991; 14(6):540–545. doi:10.1093/sleep/14.6.540

Joshua Breitstein Psy. D., Brandon Penix D. O., Bernard J. Roth M. D., Tristin Baxter AAS, Vincent Mysliwiec M. D., & F. A. A. S. M. Intensive sleep deprivation and cognitive behavioral therapy for pharmacotherapy refractory insomnia in a hospitalized patient. *Journal of Clinical Sleep Medicine*. 2014; 10:689-690. doi:10.5664/jcsm.3804

Kang, T., & Chen, T. T. Performance of the generalized S-X^2 item fit index for polytomous IRT models. *Journal of Educational Measurement*. 2008; 4:383-395. doi:10.2307/20461906

Kirisci, L., Tarter, R., Reynolds, M., Ridenour, T., Stone, C., & Vanyukov, M. Computer adaptive testing of liability to addiction: Identifying individuals at risk. *Drug and Alcohol Dependence*. 2012; 123:S79–S86. doi:10.1016/j.drugalcdep.2012.01.016

Kocalevent, R. D., Rose, M., Becker, J., Walter, O. B., Fliege, H., Bjorner, J. B., … Klapp, B. F. An evaluation of patient-reported outcomes found computerized adaptive testing was efficient in assessing stress perception. *Journal of Clinical Epidemiology*. 2009; 62(3):278–287. doi:10.1016/j.jclinepi.2008.03.003

Lange, T., Dimitrov, S., & Born, J. Effects of sleep and circadian rhythm on the human immune system. *Annals of the New York Academy of Sciences*. 2010; 1193(1):48–59. doi:10.1111/j.1749-6632.2009.05300.x

Li, J. M. the Self-Rating Scale of Sleep (SRSS). *China Journal of Health Psychology*. 2012; 12:1851-1851.

Magis, D., & Raiche, G. Random generation of response patterns under computerized adaptive testing with the R package catR. *Journal of Statistic Software*. 2012; 48:1-31. doi:10.18637/jss.v048.i08

Magis, D., & Barrada, J. R. Computerized Adaptive Testing with R: recent updates of the package catR. *Journal of Statistic Software*. 2017; 76:1-19. doi:10.18637/jss.v076.c01

Maris, E. Estimating multiple classification latent class models. *Psychometrika*. 1999; 64(2):187–212. doi:10.1007/bf02294535

Meijer, R. R., & Nering, M. L. Computerized Adaptive Testing: Overview and Introduction. *Applied Psychological Measurement*. 1999; 23(3):187–194. doi:10.1177/01466219922031310

Morin, C. M. *Insomnia: Psychological Assessment and Management*. Guilford Press; 1993.

Muraki, E. A Generalized Partial Credit Model: Application of an EM Algorithm. *Applied Psychological Measurement*. 1992; 16(2):159–176. doi:10.1177/014662169201600206

O'Connor, B. P., Crawford, M. R., & Holder, M. D. An Item Response Theory Analysis of the Subjective Happiness Scale. *Social Indicators Research*. 2014; 124(1):249–258. doi:10.1007/s11205-014-0773-9

Perez, L., Huang, J., Jansky, L., Nowinski, C., Victorson, D., Peterman, A., & Cella, D. Using Focus Groups to Inform the Neuro-QOL Measurement Tool: Exploring Patient-Centered, Health-Related Quality of Life Concepts Across Neurological Conditions. *Journal of Neuroscience Nursing*. 2007; 39(6):342–353. doi:10.1097/01376517-200712000-00005

Reckase, M. D. Unifactor Latent Trait Models Applied to Multifactor Tests: Results and Implications. *Journal of Educational Statistics*. 1979; 4 (3):65–75. doi:10.2307/1164671

Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., Cella, D. Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the patient-reported outcomes measurement information system (PROMIS). *Medical Care*. 2007; 45:S22–S31. doi:10.1097/01.mlr.0000250483.85507.04

Reise, S. P., & Waller, N. G. Item Response Theory and Clinical Measurement. *Annual Review of Clinical Psychology*. 2009; 5(1):27–48. doi:10.1146/annurev.clinpsy.032408.153553

Samejima, F. Estimation of latent ability using a response pattern of graded scores. *ETS Research Bulletin Series*. 1968; i–169. doi:10.1002/j.2333-8504.1968.tb00153.x

Samejima, F. Estimation of reliability coefficients using the test information function and its modification. *Applied Psychological Measurement*. 1994; 18:229-244. doi:10.1177/014662169401800304

Schisterman, E. F., Perkins, N. J., Liu, A., & Bondell, H. Optimal Cut-point and Its Corresponding Youden Index to Discriminate Individuals Using Pooled Blood Samples. *Epidemiology*. 2005; 16(1):73–81. doi:10.1097/01.ede.0000147512.81966

Schwarz, G. Estimating the Dimension of a Model. *The Annals of Statistics* 1978; 6(2):461–464. doi:10.1214/aos/1176344136

Smits, N., Cuijpers, P., Beekman, A. T. F., & Smit, J. H. Reducing the length of mental health instruments through structurally incomplete designs. *International Journal of Methods in Psychiatric Research*. 2007; 16(3):150–160. doi:10.1002/mpr.223

Smits, N., Cuijpers, P., & van Straten, A. Applying computerized adaptive testing to the CES-D scale: A simulation study. *Psychiatry Research*. 2011; 188(1):147–155. doi:10.1016/j.psychres.2010.12.001

Soldatos, C. R., Dikeos, D. G., & Paparrigopoulos, T. J. Athens Insomnia Scale: validation of an instrument based on ICD-10 criteria. *Journal of Psychosomatic Research*. 2000; 48(6):555–560. doi:10.1016/s0022-3999(00)00095-7

Spiegelhalter, D. J., Best, N. G., & Carlin, B. P. Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models. *Research. Report*. 1998; 98-109.

Spoormaker, V. I., Verbeek, I., van den Bout, J., & Klip, E. C. Initial Validation of the SLEEP-50 Questionnaire. *Behavioral Sleep Medicine*. 2005; 3(4):227–246. doi:10.1207/s15402010bsm0304_4

Stocking, M. L., Swanson, L. A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement*. 1993; 17:277- 292. doi:10.1177/014662169301700308

Swets, J. A. Measuring the accuracy of diagnostic systems. *Science*. 1998; 240:1285–93. doi:10.1126/science.3287615

Tan, Q., Cai, Y., Li, Q., Zhang, Y., & Tu, D. Development and Validation of an Item Bank for Depression Screening in the Chinese Population Using Computer Adaptive Testing: A Simulation Study. *Frontiers in Psychology*. 2018; 9. doi:10.3389/fpsyg.2018.01225

Wu, M. 300 million Chinese suffer from sleep disorders. *The march wind*. 2019; 4:7-7.

Yao, Y. M., Jiang, C. Q., Ma, X., Xi, Y. J., Tian, L., He, Y. Y., Zhang, H., & Yang, F. C. Content validity test of Beijing residents mental health scale. *China Journal of Health Psychology*. 2017; 25(6):873-877.

Zai, Q., Hou, J. C., & Fang, X. M. Sleep disorders have an influence on immune function and the development of sepsis. *International Journal of Anesthesiology and Resuscitation*. 2017; 38(10):934-937. doi:10.3760/cma.j.issn.1673-4378.2017.10.016

Zhang, L. Y., Kong, L. M., Zhang, Q. J., Tao, F. Y., Ma, A. G., Liu, Y., Gao, Y. F., Tu D. H., Su, W. J., & Wang L. J. Development and validity test of Chinese sleep disorder scale. *World Journal of Sleep Medicine*. 2014; 3:140-146.

Zhang, H. J., Zhao, Z. X. Main progress on the clinical research of sleep disorders. *Journal of Neuroscience and Mental Health*. 2015; 15(1):6-8. doi:10.3969/j.issn.1009-6574.2015.01.002

Figures

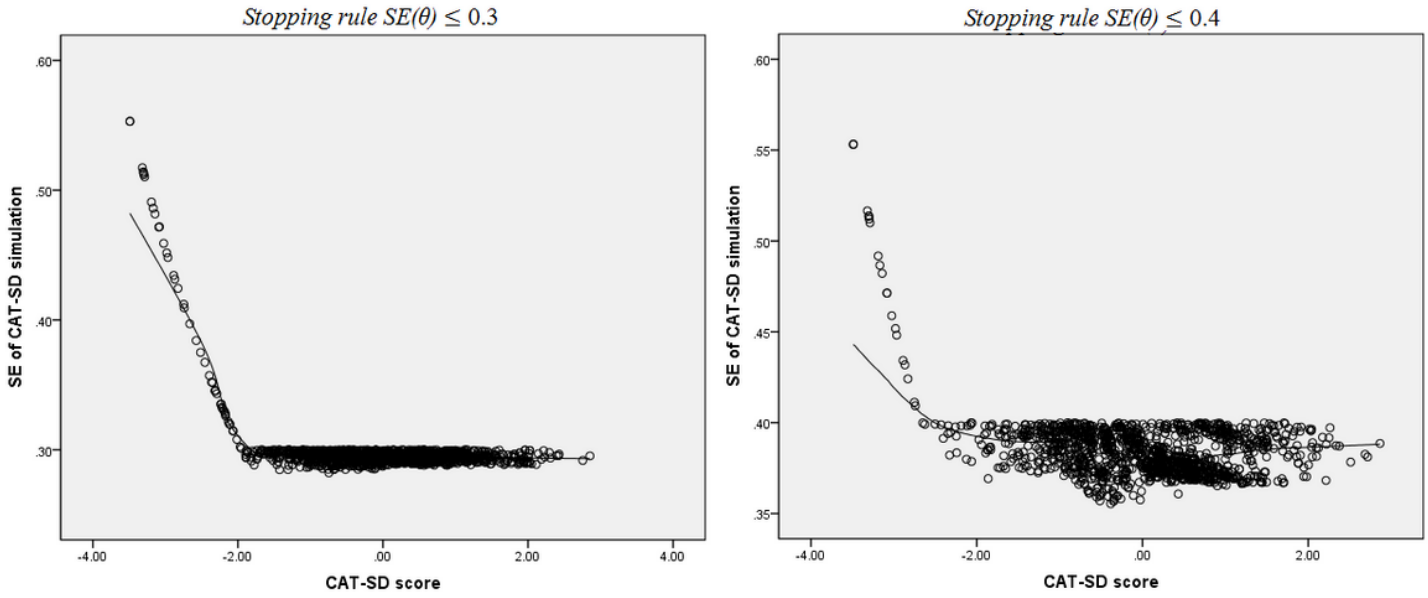

Figure 1

Standard error (SE) of CAT-SD score under stopping rules SE(θ) ≤ 0.3 and SE(θ) ≤ 0.4. Note: A plot suggests good measure precision for the majority of the CAT-SD score range.

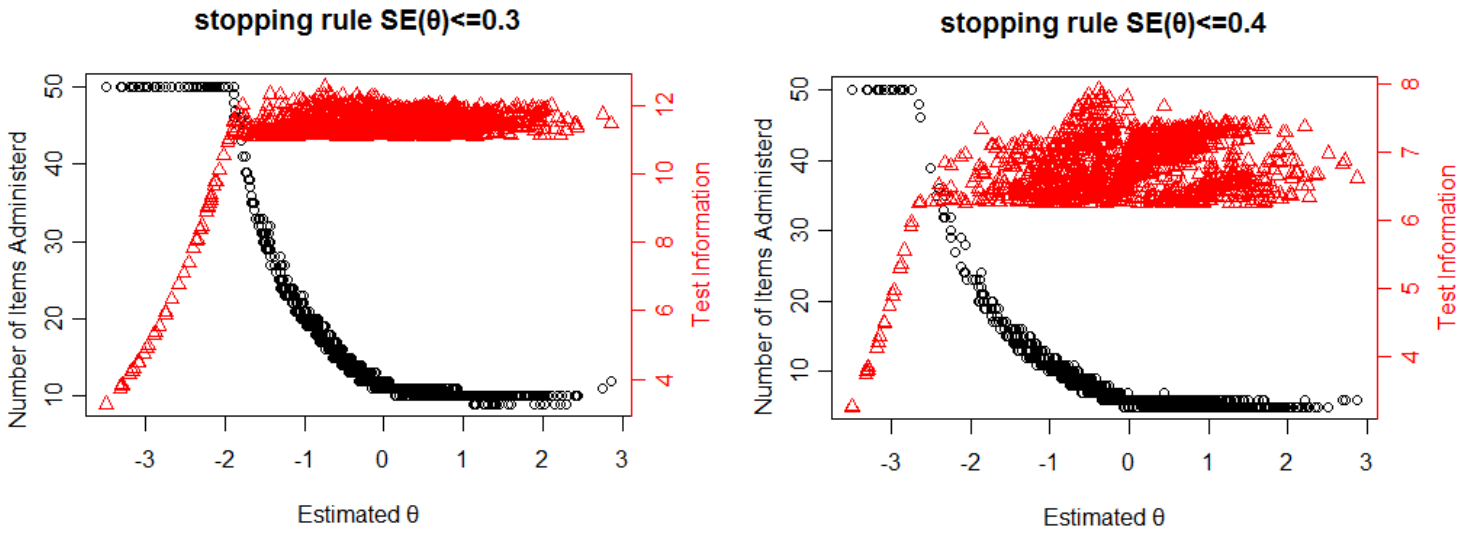

Figure 2

Number of selected items and test information curve under stopping rules SE(θ) ≤ 0.3 and SE(θ) ≤ 0.4.

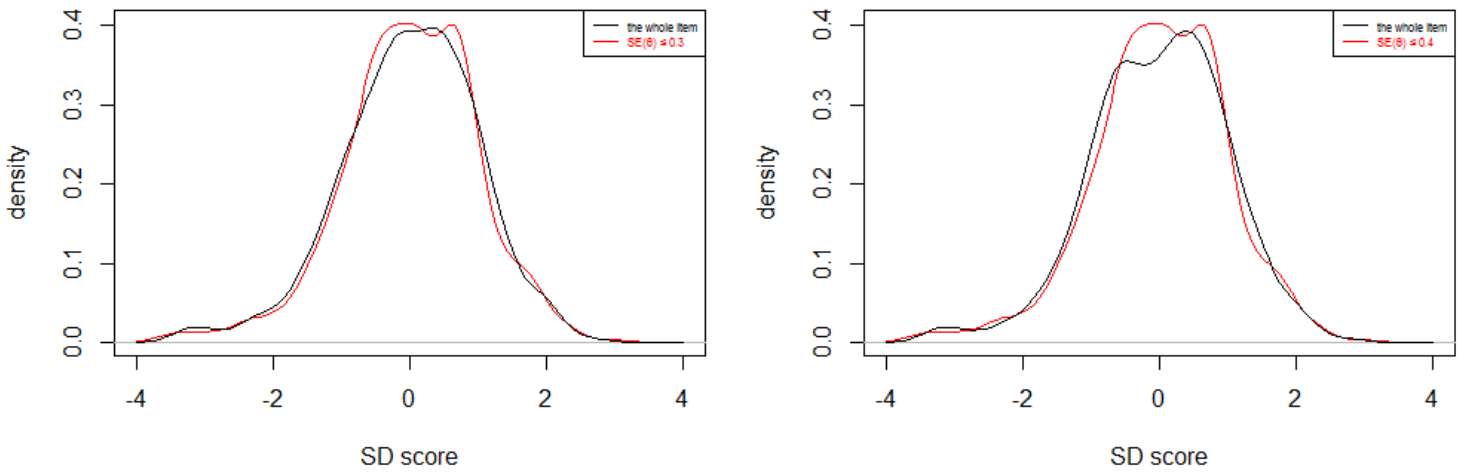

Figure 3

Density distribution of SD score between CAT-SD (stopping rules SE(θ) ≤ 0.3 and SE(θ) ≤ 0.4) and the whole item bank.

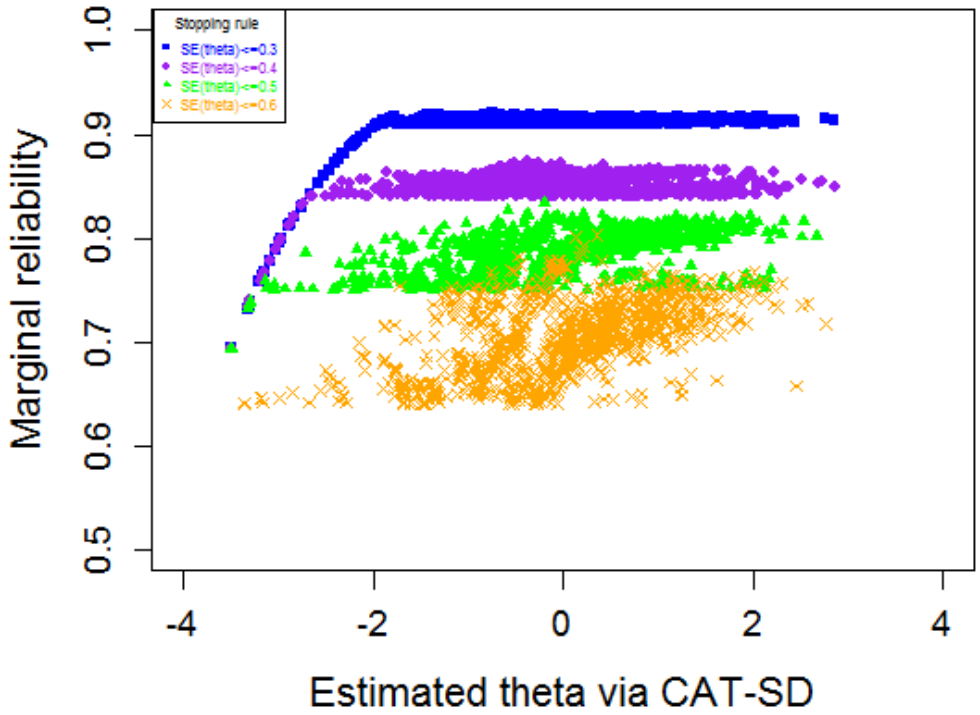

Figure 4

Marginal reliability as a function of estimated theta under different stopping rules.

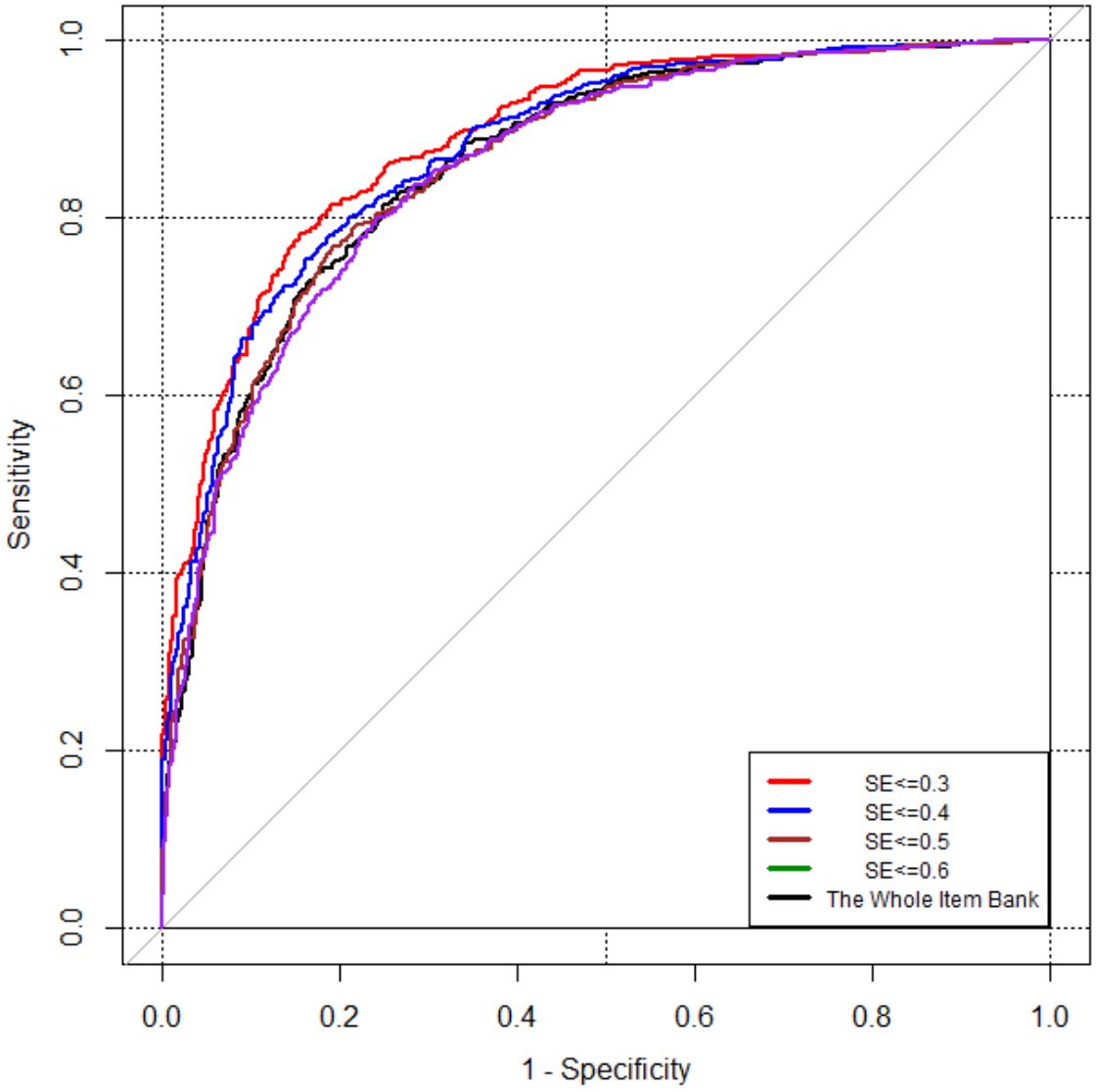

Figure 5

The ROC curve of CAT-SD with several stopping rules.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.pdf