ELSEVIER

Sleep Medicine 10 (2009) 556–565

SLEEP MEDICINE

www.elsevier.com/locate/sleep

Original Article

Development of a Japanese version of the Epworth Sleepiness Scale (JESS) based on Item Response Theory ☆

Misa Takegami [a,*], Yoshimi Suzukamo [b], Takafumi Wakita [a], Hiroyuki Noguchi [c], Kazuo Chin [d], Hiroshi Kadotani [e], Yuichi Inoue [f], Yasunori Oka [f], Takaya Nakamura [d], Joseph Green [g], Murray W. Johns [h], Shunichi Fukuhara [a]

[a] *Department of Epidemiology and Healthcare Research, Graduate School of Medicine and Public Health, Kyoto University, Yoshida konoe-cho, Sakyo-ku, Kyoto 606-8501, Japan*

[b] *Department of Physical Medicine and Rehabilitation, Graduate School of Medicine, Tohoku University, Sendai, Japan*

[c] *Department of Psychology and Human Development Sciences, Graduate School of Education and Human Development, Nagoya University, Aichi, Japan*

[d] *Department of Respiratory Medicine, Graduate School of Medicine, Kyoto University, Kyoto, Japan*

[e] *Center for Genomic Medicine, Graduate School of Medicine, Kyoto University, Kyoto, Japan*

[f] *Japan Somnology Center, Neuropsychiatric Research Institute, Tokyo, Japan*

[g] *Graduate School of Medicine, University of Tokyo, Tokyo, Japan*

[h] *Sleep Diagnostics Pty Ltd., Melbourne, Australia*

Received 28 February 2008; received in revised form 28 April 2008; accepted 28 April 2008
Available online 27 September 2008

Abstract

Background: Various Japanese versions of the Epworth Sleepiness Scale (ESS) have been used, but none was developed via standard procedures. Here we report on the construction and testing of the developer-authorized Japanese version of the ESS (JESS).

Methods: Developing the JESS involved translations, back translations, a pilot study, and psychometric testing. We identified questions in the ESS that were difficult to answer or were inappropriate in Japan, proposed possible replacements for those questions, and tested them with analyses based on item response theory (IRT) and classical test theory. The subjects were healthy people and patients with narcolepsy, idiopathic hypersomnia, or obstructive sleep apnea syndrome.

Results: We identified two of our proposed questions as appropriate replacements for two problematic questions in the ESS. The JESS had very few missing data. Internal consistency reliability and test–retest reliability were high. The patients had significantly higher JESS scores than did the healthy people, and higher JESS scores were associated with worse daytime function, as measured with the Pittsburgh Sleep Quality Index.

Conclusions: In Japan, the JESS provides reliable and valid information on daytime sleepiness. Researchers who use the ESS with other populations should combine their knowledge of local conditions with the results of psychometric tests.

Keywords: Epworth Sleepiness Scale; Daytime sleepiness; Item Response Theory; Psychometric propensities; Validity; Reliability

1. Introduction

Daytime sleepiness is an important manifestation of sleep disorders. It can disrupt a patient's social life and threaten public health and safety [1–4]. Daytime sleepiness is an important marker in assessments of sleep

☆ *Disclosure statement:* This study was supported by grants from the Institute for Health Outcomes and Process Evaluation Research (iHope International). All authors have indicated no financial conflict of interest.

* Corresponding author. Tel.: +81 75 753 4645; fax: +81 75 753 4644.

E-mail address: takegami-kyt@umin.ac.jp (M. Takegami).

1389-9457/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.sleep.2008.04.015

disorders, and is measured both subjectively and objectively. The "gold standard" index of sleepiness is provided by the Multiple Sleep Latency Test (MSLT) [5], but this test is costly and time-consuming. Requiring much less time and money is the Epworth Sleepiness Scale (ESS), a self-report instrument for measuring a patient's perception of sleepiness. Guidelines for the control of Obstructive Sleep Apnea Syndrome (OSAS), narcolepsy, and insomnia recommend the ESS [6–8], and it has also been used in occupational and community-based studies [9,10].

The ESS comprises questions about subjective sleepiness in eight situations [11,12]. Respondents use a 4-point scale (scored 0–3) to respond to each of the eight questions, and the scores are summed to give an overall score of 0–24. Higher scores indicate stronger subjective daytime sleepiness, and scores below 10 are considered to indicate no problem [9].

Several Japanese language versions of the ESS are available, but their relation to the original version and their acceptability are questionable because none was developed in accordance with standard procedures or in coordination with the developer of the original (English language). An advisory committee on sleep apnea syndrome within the Japanese Respiratory Society reported in 2004 that 165 of 277 hospitals used various Japanese language versions of the ESS, and used them in different ways (self-administration or interview). This inconsistency hampers comparisons among hospitals [13]. In addition, although many papers published from Japan, including one by us [14], have reported ESS data, the possibility remains that the sleepiness measured with those versions differs in important ways from that measured with the original ESS. Furthermore, the questions in the original ESS ask about sleepiness in various daily life situations, but we should not assume that all of those situations are familiar to respondents in Japan. Such a lack of familiarity could explain the reported high rates of missing data (9.5–19.2%) [15]. For example, the ubiquity of public transportation in Japan could account for the fact that 19.2% of the subjects in one study did not answer question 8, which asks about sleepiness "In a car, while stopped for a few minutes."

Item Response Theory (IRT) is increasingly used in the construction of scales for measuring subjective attributes, particularly in research on Quality of Life. Common applications of IRT are the construction of shorter versions of existing scales, establishment of scoring algorithms, and Computerized Adaptive Testing (CAT) [16–19]. In IRT, the probability of a particular response to a question is described as a function of a latent trait assumed to underly the manifest response. In IRT the value of the latent trait is called "theta" [20,21], and in this study it is the actual subjective daytime sleepiness. The probability of a particular response is typically described with a function that has two parameters: "location" is the value of the latent trait about which a response provides the most information (in educational testing this is called "difficulty"), and "slope" is the degree to which responses can be used to distinguish between small differences in the latent trait [22]. In the ESS, questions with higher values of the location parameter provide more information about people whose daytime sleepiness is severe, and questions with lower values of the location parameter provide more information about people whose daytime sleepiness is milder. Questions with higher values of the slope parameter allow one to make fine distinctions between severities of daytime sleepiness, i.e., to measure small differences in daytime sleepiness, while questions with lower values of the slope parameter allow one to measure only relatively larger differences in daytime sleepiness. Analyses based on IRT allow more precise examinations of the characteristics of each question item than do those based on Classical Test Theory (CTT) [20,21]. In addition to analyses of each question, IRT allows construction of an information function for each question and also for the scale as a whole. That test information function reflects the accuracy of measurement at different values of the latent trait [20,21]. Given a sufficiently large and wide-ranging group of questions (the "item pool"), and knowledge of each question's location and slope, a scale with a desired test information function can be constructed.

Here, at the request of the Japanese Respiratory Society, we report the development of a Japanese version of the ESS (the JESS). We were able to ensure that the JESS was as close as practically possible to the ESS, because one of us (M.J.) is the developer of the ESS. Our purposes were to translate the ESS into Japanese using commonly-accepted methods, to clarify problems with unauthorized Japanese-language versions of the ESS, to use IRT to develop a better Japanese-language version (the JESS), and to study the reliability and validity of the JESS.

2. Methods

2.1. Translation and preparation of the item pool

To develop the Japanese version we used a method that has been used in many countries and for many self-report scales [23]. The process includes translation from the source language into the target language (i.e., forward translation), translation from the target language back into the source language (i.e., back translation) so the developer of the source-language version can participate fully, and examination of translation quality. We also included a pilot test.

At the forward translation step, two types of equivalence were examined: (1) conceptual equivalence, i.e., whether the two versions of the scale can be expected to measure the same theoretical construct; and (2)

technical equivalence, i.e., whether the methods of assessment are comparable with respect to data collection [24]. Two translators, both of whom are native speakers of Japanese who have a good command of English, translated the ESS into Japanese separately. Members of the research team then discussed differences between the two translations and integrated them into one that was easily understandable while maintaining semantic equivalence (equivalence regarding the meaning of the terms) with the ESS [24]. We also examined whether the content of each question was relevant to the lives and experiences of the intended respondents. To replace problematic questions, i.e., questions for which data were frequently missing in a community-based survey [15], we wrote new questions. We formed a group that included a psychologist, a clinical epidemiologist, physicians, and sleep researchers. Through its discussions, the group decided on questions to be tested as potential replacements for the original problematic questions. The group used two criteria to make those decisions: (1) the questions should describe situations that the group believes are commonly encountered in the daily lives of Japanese respondents, and (2) the questions should describe situations that the group believes carry the same potential for dozing off as was intended by the original questions (i.e., they should be expected to have a similar value of theta). The selected potential replacement items were reviewed by the original developer (M.J.) and by a specialist in sleep medicine in Japan.

The back translation was done by a translator who was a native speaker of English and had a good command of Japanese. The back translated version was reviewed by the original developer of the ESS, and differences between that version and the original version were discussed, particularly with regard to differences in lifestyle between the intended respondents to the two versions. Those of us in Japan then discussed the problems pointed out by the original developer, modified the forward translation accordingly, had the modifications back translated, and consulted the original developer again.

For the pilot test, 10 healthy people responded to the questions in the modified forward translation (and to the new questions), after which structured interviews were conducted with each respondent individually. To find out whether there were any unanticipated translation errors, the interviewer asked the respondent to restate or interpret each question in the respondent's own words. The interviewer asked whether the question had any unnecessarily complex or sophisticated wording, whether it was easy or difficult to comprehend, and whether the respondent would have asked the question in a different way. The interviewer also asked whether they regularly encountered the situations described in the new questions, and whether they could imagine themselves in those situations even if they do not regularly encounter them. We repeated these steps until the developer authorized a final version of the JESS.

2.2. Subjects and data collection

Cross-sectional data were obtained from a convenience sample of 143 outpatients and 397 healthy subjects. They completed a self-administered questionnaire between October 2004 and December 2005. The outpatients had Obstructive Sleep Apnea Syndrome (OSAS), narcolepsy, or idiopathic hypersomnia, and had sought treatment at either of two outpatient hospitals specializing in sleep disorders. For the healthy subjects, we used personal networks to recruit a convenience sample of people who were not being treated for a sleep problem: employed men, undergraduate students, and retirees.

The questionnaire included the translated ESS (both the original questions and the new questions being tested as potential replacements for problematic original questions), and also questions about date of birth, medical history, sex, height, weight, sleep duration, the presence or absence of snoring, and whether the subject possessed a driver's license. For the outpatients, clinicians provided information on the degree of severity of the sleep disorder before treatment, treatment method and period, blood pressure, and physiological indexes related to sleep disorders (apnea–hypopnea index, oxygen desaturation index, and minimum SaO_2).

To examine test–retest reliability, from the healthy subjects, 57 employed men completed a questionnaire (with the translated ESS only) 1 week after the first data collection. To examine responsiveness, 32 OSAS patients who had not undergone treatment with Continuous Positive Airway Pressure (CPAP) at the time of the first data collection were studied again 1 month after treatment with CPAP began.

Informed consent was obtained, and the study was approved by the Institutional Review Board of Kyoto University, Graduate School of Medicine and by the Institutional Review Board of the Neuropsychiatric Research Institute.

2.3. Data analysis

We used data from 540 respondents. We divided the respondents randomly into a development sample and a validation sample ($n = 270$ each). First, data from the development sample were used in IRT-based analyses. Then the results of those analyses were used to select new questions to replace problematic original questions. Finally, data from the validation sample were used in analyses based on CTT, to evaluate the psychometric properties of the JESS. We used Parscale 4.1 (SSI,

Inc., 2003) and SPSS12.0J for Windows (SPSS, Inc., 2003).

2.4. IRT-based analyses

A precondition for analysis based on IRT is unidimensionality of the item pool (this refers to the original questions considered together with the new questions being tested as potential replacements for problematic original questions). To find out whether the item pool was unidimensional, we used factor analysis and examined the pattern of successive eigenvalues (scree test) [25]. The Generalized Partial Credit Model (GPCM) [22] was used to analyze ordered response categories. The GPCM estimates three parameters: the slope parameter, the location parameter, and the category parameter. The category parameter corresponds to each response category and, in conjunction with the location parameter, it reflects theta [20,21]. In the GPCM, the category parameter is assumed to be the same for all questions with the same response categories. The item parameters were estimated using marginal maximum likelihood estimation [26]. A potential replacement question was considered to be a good replacement if, when compared with a problematic original question, it had a similar location but a greater slope, i.e., if it gave information about a similar severity of daytime sleepiness, but did so with greater precision. To illustrate that psychometric improvement offered by the replacement questions, we drew Item Response Category Characteristic Curves, which show response probability as a function of theta.

We also examined the possibility that responses to question 8 depended, at least to some extent, not on sleepiness but on the respondent's experience as an automobile driver, which would be an instance of "differential item functioning" [27]. Specifically, we tested the hypothesis that having a driver's license was associated with the responses to question 8, but not with the responses to question 13, because question 13 was not related to automobiles but was at almost the same location as question 8.

For an overall comparison of the scales before and after the substitution of problematic questions with new questions, we drew test information functions for the two scales [20,21]. As their name implies, these functions show how much information the results can be expected to provide, as a function of theta.

2.5. Psychometric properties of the JESS

2.5.1. Item analysis

The percentage of missing values was computed for each question. Means and standard deviations of the JESS total scores were computed. Questions to which 90% or more of the subjects gave the same response were also noted.

2.5.2. Reliability

Cronbach's alpha was used as the index of internal consistency reliability [28]. Values of about 0.7 or higher are generally accepted as evidence of good internal consistency reliability. Test–retest reliability (1-week interval) was assessed using the intra-class correlation coefficient (ICC) [29].

2.5.3. Validity

We hypothesized that the JESS measures only one latent trait (subjective daytime sleepiness), and used factor analysis to test the hypothesis of unidimensionality.

Testing for concurrent validity, we examined the association between JESS scores and scores from the "daytime dysfunction" domain of the Pittsburg Sleep Quality Index (PSQI) [30,31]. The PSQI has been translated into Japanese and tested for validity [32,33]. In the PSQI, daytime dysfunction refers to, for example, difficulty staying awake during social activities and lack of sufficient motivation to get things done. The PSQI is scored with a 4-point scale (scores from 0 to 3), with higher scores indicating stronger daytime dysfunction.

To test for known-groups validity we compared patients with healthy people. We used analysis of covariance to estimate the adjusted mean differences between groups, estimated least-squares means of JESS scores with adjustments for age and sex, and used Bonferroni's correction for multiple comparisons.

2.5.4. Responsiveness

Responsiveness was examined using data from OSAS patients who completed the survey at baseline and again one month after the start of CPAP treatment. Differences associated with the treatment were analyzed with Student's *t*-test for paired data.

3. Results

3.1. Translation and item pool

Discussions between the Japanese team and the developer of the original version resulted in some differences between the JESS and previous Japanese-language versions.

In our first translation and in several previous Japanese-language versions the instructions and response options used the expression "nemutteshimau," which means "fall asleep." Because "dozing off" was meant to indicate sleeping for a short time, we instead used "utoutosuru (suubyou~suufun nemutteshimau)," which means "doze off (fall asleep for several seconds to several minutes)." Another important point is that several previous Japanese language versions ask about the frequency of falling asleep, but the ESS asks, with regard to the eight situations, not about frequency but about likelihood. It asks not *how often* one dozes off but *how*

likely one is to doze off. We decided that the question sentence and the response options needed to be phrased in terms of the possibility of dozing off rather than the fact of dozing off, and so we changed them to "How great is the possibility of you dozing off?" and "There is no possibility of dozing off" and so on. We also added a postscript, namely that "It is very important that you answer all items. Please try, as much as possible, to answer all of the items." These changes may have contributed to the relatively low rate of missing data, which is described in the section on item analysis below.

According to the original developer, question 8 ("in a car while stopped for a few minutes") includes being in a car not only as the driver but also as a passenger. However, in our first translation and in several previous Japanese language versions, question 8 had been mistranslated as "while driving" or "when driving a car." In fact, even the original English version of question 8 has been misinterpreted with regard to its reference to being in a car as a driver, as a passenger, or either one (personal communication, MJ). Consequently, we decided to replace question 8 with one involving a different situation and activity. The developer approved six new questions as potential replacements (questions 10 through 15 in Table 2).

The situation described in question 1 was translated in one of our preliminary versions and in other previous Japanese language versions as "while sitting and reading a book." In one study [15], almost 15% of the subjects did not respond to that question. We thought that at least some of the respondents commonly read newspapers or magazines rather than books, so we included question 9 (Table 2) as a possible replacement for question 1.

In the pilot study, the respondents indicated that they clearly understood the difference between questions about the *frequency* of falling asleep and questions about the *possibility* of falling asleep. They understood the former as asking about their actual experiences in daily life, and the latter as asking them to guess how sleepy they would be in a given situation. In the latter case they could answer questions about situations that they do not regularly experience. None of the respondents said that questions about the possibility of falling asleep were difficult to answer. The respondents also indicated that they regularly encountered many of the situations described in the new questions, and they said they could imagine how they would respond in those situations that they do not regularly encounter.

3.2. Subject characteristics

Characteristics of the subjects are shown in Table 1 ($n = 540$). The healthy subjects were 229 male salaried employees, 125 undergraduate students, and 43 retirees ($n = 397$; 75.9% men; mean age 38.9, SD = 14.7). Among the patients were 85 with OSAS, 54 with narcolepsy, and four with idiopathic hypersomnia ($n = 143$; 70.2% men; mean age 47.9, SD = 18.1).

3.3. Item selection using IRT

The successive eigenvalues (scree test) from factor analysis were 5.77, 1.42, 1.25, 1.11, and 0.81. The pattern in that series, specifically the large difference between the first two eigenvalues together with the relatively small differences among further eigenvalues, is evidence in favor of a one-factor solution. The factor loadings ranged from 0.47 to 0.76, which is further evidence of unidimensionality of the JESS.

As shown in Table 2 (the GPCM results) and in Fig. 1, the location parameters of questions 1 and 9 were similar and the curves are in approximately the same place on the graph with regard to theta. However, the slope parameter of question 9 was greater than that of question 1 and the curves for question 9 are steeper than those for question 1. Therefore, we replaced question 1 with question 9.

Similarly, questions 8 and 13 were in approximately the same location, but question 13 had greater slope(s). Furthermore, the location parameter of question 8 differed greatly between subjects who had a driver's license and those who did not: 2.28 (SE = 0.20) vs. 3.16 (SE = 0.45). In contrast, the location parameter of question 13 differed much less between those subjects who did and those who did not have a license: 2.18 (SE = 0.16) vs. 2.00 (SE = 0.24). For those reasons we replaced question 8 with question 13.

The JESS, i.e., the version with questions 9 and 13 instead of questions 1 and 8, provided more information

Table 1
Subject characteristics

	Development sample $n = 270$	Validation sample $n = 270$
Men, %	73.3	73.0
Age, mean (min–max)	42.1 (14–82)	40.6 (16–79)
Body mass index, mean (standard deviation)	23.9 (4.5)	23.9 (4.5)
Snoring everyday or often, %	38.9	38.5
Sleep medication use, %	12.2	8.5
Smoking, %	32.2	32.2
Drinking almost every day, %	36.6	34.4
Possess driver's license, %	80.0	82.2
Disease		
Obstructive sleep apnea syndrome before treatment, %	10.7	8.9
Obstructive sleep apnea syndrome during treatment, %	4.8	7.0
Narcolepsy before treatment, %	2.6	1.5
Narcolepsy during treatment, %	6.7	9.3
Idiopathic hypersomnia before treatment, %	1.1	0.4

Table 2
Item characteristics estimated with analyses based on IRT in the development sample ($n = 270$)

		a	SE	b	SE
Original question-items					
1	While sitting and reading a book	0.69	0.07	0.46	0.08
2	While watching a television	0.79	0.08	0.68	0.07
3	While sitting at a meeting or a theater without actively speaking, etc	0.62	0.07	0.25	0.08
4	While being in a car as a passenger for one consecutive hour	0.55	0.06	-0.00	0.08
5	While lying down and taking a rest in the afternoon	0.75	0.09	-0.96	0.08
6	While sitting and talking with someone	0.76	0.11	2.36	0.16
7	While sitting quietly after taking lunch (without liquor)	0.83	0.09	0.06	0.06
8	In a car while stopped for a few minutes because of a signal or a traffic jam	0.58	0.09	2.41	0.21
Additional question-items					
9	Reading something while sitting in a chair (newspapers, magazines, books, documents, etc)	1.11	0.11	0.70	0.06
10	While eating with your family or acquaintances	0.89	0.16	3.05	0.23
11	While sitting in a train, bus, or a taxi without talking to anyone	0.78	0.08	0.05	0.07
12	While soaking in the bathtub during a bath	0.39	0.06	1.88	0.20
13	While sitting and writing by hand	0.70	0.10	2.13	0.15
14	While eating alone, seated	0.82	0.14	2.91	0.22
15	While sitting and waiting for a train or a bus without talking to anyone	0.52	0.07	1.81	0.16

a, slope parameter; b, location parameter; SE, standard error. Category parameter: 1.28, −0.36, −0.92.

than the Japanese version with the original questions (Fig. 2). In particular, more information was available at higher levels of daytime sleepiness.

3.4. Psychometric properties of the JESS

3.4.1. Item analysis

For the JESS, the percentages of missing values ranged from 0.7% to 1.1%. For the original question asking about being in a car 4.4% of the data were missing, but for the new question asking about sitting and writing only 0.7% were missing. For the new question asking about reading while in a chair the percentage of missing data was also 0.7%. None of the questions had 90% or more responses in one response category. The mean JESS total score was 8.9 (SD = 4.8).

3.4.2. Reliability

Cronbach's alpha coefficient for the JESS total score was 0.85. The alpha coefficients were calculated for each 7-question scale made by eliminating one of the 8 questions; those alpha coefficients ranged from 0.82 to 0.84, which indicates that none of the questions had an unusually strong influence on the internal consistency. Test–retest reliability was high (intraclass correlation coefficient = 0.78, $n = 30$).

3.4.3. Validation testing

Item-total correlation coefficients ranged from 0.53 to 0.69. Factor analyses of data from the patients and the healthy subjects indicated that the JESS is unidimensional in both groups. Loadings on the first factor ranged from 0.58 to 0.77 (Table 3).

Higher JESS scores were associated with worse daytime function (as measured with the PSQI, Fig. 3, ANCOVA p-value for trend < 0.001). The JESS total score was 6.8 for the group with the best daytime function and 13.6 for the group with the worst daytime function.

The patients had significantly higher JESS scores than did the healthy people (10.8 vs. 8.4, $p < 0.001$).

3.4.4. Responsiveness

Fourteen patients with OSAS were randomly allocated to the validation sample, and complete ESS data were obtained from 12 of them. In those 12 patients, CPAP treatment was associated with a 3.67-point improvement in JESS scores ($p = 0.007$).

4. Discussion

This study showed that the authorized Japanese translation of the ESS (the JESS) measures a construct similar to that measured by the original ESS. Moreover, we showed how the original ESS question about being in a car was problematic, and we identified an appropriate replacement for that question.

While the original ESS asked questions about probabilities ("how likely"), previous Japanese versions asked factual questions ("how often") and measured more severe sleepiness. To avoid that content inequivalence and other potential problems, we used forward translations, back translations, examinations of translation quality, consultations with the developer of the original ESS, and a pilot test.

The most problematic question in the original ESS was the one asking about sleepiness while in a car. The percentage of missing data was high, and when

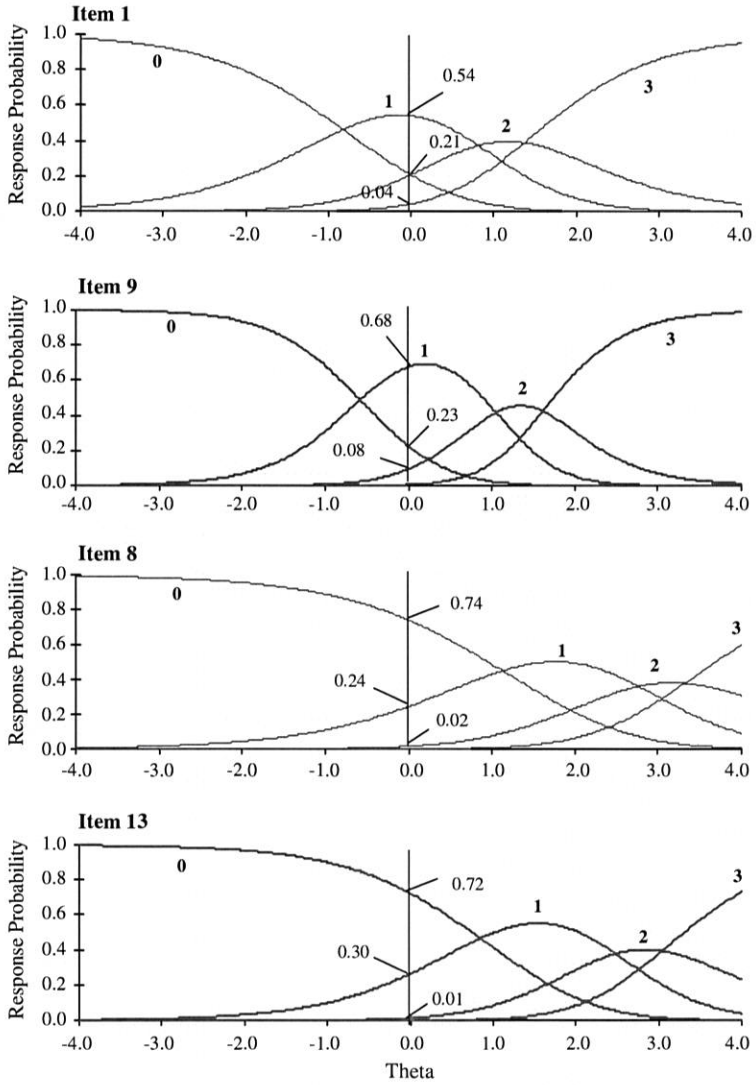

Fig. 1. Item response category characteristic curves used by the Generalized Partial Credit Model. The curves show response probability functions for each response choice. The response options were "There is no possibility of dozing off" (0), "There is a slight possibility of dozing off" (1), "There is a possibility of dozing off once in two times" (2), and "There is a high possibility of dozing off" (3). Question-item #1 was replaced with #9, and #8 was replaced with #13. Category parameter: 1.28, −0.36, −0.92.

subjects did respond to that question the responses varied between people with and without a driver's license. Using the results of analyses based on IRT, we replaced that question with an equivalent question about sleepiness in a situation that is more familiar to many Japanese people: sitting and writing.

Although being "in a car while stopped for a few minutes because of a signal or a traffic jam" might seem to be very different from "sitting and writing (by hand)," the questions about sleepiness in those two situations had similar values of the location parameter, i.e., they were close to each other on the dimension interpreted as daytime sleepiness. In addition, responses to the question about sitting and writing were more sensitive to small differences in sleepiness. In the JESS, the question about sitting and writing had the second highest value of the location parameter, which is consistent with results of a previous study using the original English version of the question about sleepiness while "in a car while stopped for a few minutes" [34].

In some other countries, the question about "sitting and reading a book" (question 1) had the sixth highest value of the location parameter [34,35]. However, in the JESS the replacement for that question (question 9 in Table 2) had the third highest value. The level of daytime sleepiness associated with two outwardly similar activities may vary with differences in habits in and socio-cultural context.

Replacing question 1 with question 9 and question 8 with question 13 resulted in a scale that can provide

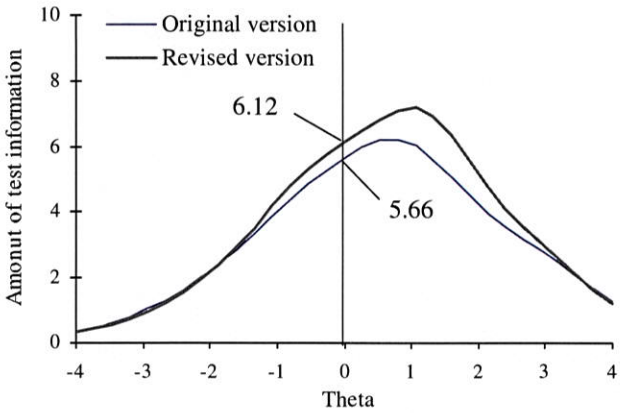

Fig. 2. Test information curves of the original and revised versions. "Original version" means the JESS with eight original questions, and "revised version" means the JESS with the two new questions used on the basis of the results of IRT analyses.

more information. Looking at the test-information function (Fig. 2), we note that its peak is not far from the right end of the axis representing theta, and that the increase in test information from the original to the revised version is also mainly on the right side, both of which suggest that the JESS will be particularly useful for patients whose daytime sleepiness is severe.

Results of psychometric tests based on CTT indicate that the JESS has good internal consistency reliability, test–retest reliability, factor validity, criterion-based validity, and responsiveness. The factor structure of the JESS is similar to that of the English version [12].

In a community-based survey that used a previous Japanese version of the ESS, the response rate was low [15], but in the present study of a convenience sample the rate of response to the JESS was high. Further research is needed to measure the response rate to the JESS in community-based surveys.

To inform our choices about which questions to include in the JESS, we used IRT rather than CTT. Using IRT allowed us to describe the characteristics of each question in greater depth than if we had used only CTT. It also allowed us to locate each question on a unidimensional scale that represents the severity of daytime sleepiness [20,21]. For the IRT-based analyses we did not distinguish healthy subjects from patients with sleep disorders, because the ESS can be used to measure daytime sleepiness both in healthy people and in patients, and also because the parameter estimates (especially the estimate of the location parameter) would not be reliable if the severity of daytime sleepiness was not very wide. IRT analysis is most useful if the range of theta (in this case, the range of severity of daytime sleepiness) is wide. Additionally, the IRT-based calculation of test information let us examine whether, and exactly how, the precision of the scale was improved by replacing question 1 with question 9 and question 8 with question 13.

The high correlation between sleepiness and traffic accidents has become an important public safety issue [2,3]. In 2002, Japanese laws were revised to add "sleep disorders with serious symptoms of drowsiness" to the conditions under which an application for a driver's license can be refused or an existing license can be revoked. The JESS can be used to evaluate the risk of daytime sleepiness, but the fact that the sleepiness it measures is self-reported raises important concerns

Table 3
Factor loadings for each item in the revised JESS for all subjects, the patients and the healthy people in the validation sample

JESS question number	All subjects $n = 270$	Patients $n = 73$	Healthy people $n = 197$
1	0.771	0.803	0.736
2	0.592	0.692	0.516
3	0.636	0.666	0.695
4	0.687	0.660	0.702
5	0.579	0.561	0.600
6	0.611	0.728	0.517
7	0.708	0.730	0.693
8	0.600	0.671	0.524

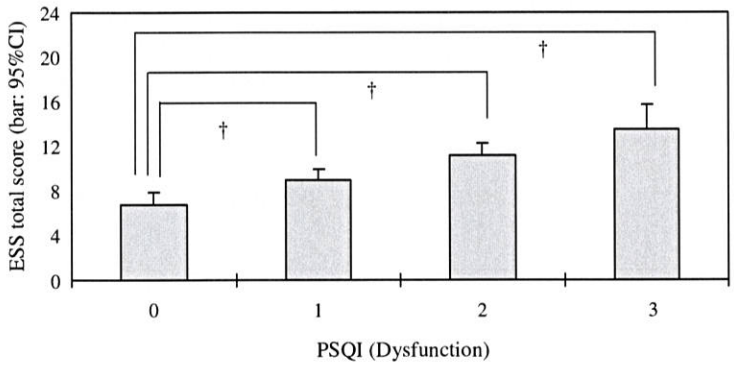

Fig. 3. Associations between JESS total score and scores on the daytime dysfunction domain of the Pittsburg sleep quality index (PSQI), after adjustment for sex and age. Daytime dysfunction domain scores of the PSQI ranged from 0 to 3, with higher scores indicating more severe daytime dysfunction. $^{\dagger}p < 0.001$, ANCOVA; p-value for trend < 0.001.

about the potential for bias, which also indicates a need for further study using IRT.

The ESS is useful only to the extent that the situations about which it asks are commonly encountered, or at least easy to envision, in daily life. To develop a version for use in Japan, merely translating the words in the original English version into the Japanese language was not sufficient; the familiarity of the situations described in the questions had to be taken into account. For example, daily life for many people in modern Japan involves very little driving or riding in cars. We found that questions 9 and 13 are good replacements for questions 1 and 8 in Japan, but those results cannot be generalized to other countries. Specifically, they cannot be generalized to groups with very different lifestyles and daily experiences. Researchers who want to similarly adapt the ESS for use with other populations should combine their knowledge of local conditions with the results of psychometric tests to decide which questions to include.

This study had three limitations. First, sleepiness in daily life may differ from sleepiness in the dark laboratory conditions used to measure the MSLT [36], and we measured only the former. Therefore, we make no claims about associations between the MSLT and the JESS. Second, we used a convenience sample in Japan. For studies of other groups, the most appropriate questions regarding daytime sleepiness might differ from those that we found to be appropriate in Japan. Users of the ESS may find the methods of the present study applicable as they adapt the ESS to the habits and experiences common among the people they wish to study.

5. Conclusion

Using standard, internationally recognized methods we developed and tested a version of the ESS for use in Japan (the JESS). To our knowledge, this study is the first application of IRT to the selection of questions for the ESS. Two questions from the original ESS were replaced. The two new questions were psychometrically similar to, or better than, the originals. The JESS is characterized by content equivalence with the original ESS and appropriateness for use in Japan, and its use is expected to result in fewer missing data than previous versions. The approach we used in this study (translations, back translations, consultations, revisions, a pilot study, IRT, and CTT) can also be used to re-examine the suitability of the ESS questions in other groups with particular common habits and experiences.

Acknowledgements

This study was supported by Grants from the Institute for Health Outcomes and Process Evaluation Research (iHope International). We are grateful to Itsunari Minami and Yuriko Nakayama for recruiting subjects and collecting data. We also wish to thank Tsutomu Namikawa for his suggestions on our analysis.

References

[1] Akashiba T, Kawahara S, Akahoshi T, Omori C, Saito O, Majima T, et al. Relationship between quality of life and mood or depression in patients with severe obstructive sleep apnea syndrome. Chest 2002;122(3):861–5.
[2] Findley LJ, Fabrizio M, Thommi G, Suratt PM. Severity of sleep apnea and automobile crashes. N Engl J Med 1989;320(13):868–9.
[3] Lyznicki JM, Doege TC, Davis RM, Williams MA. Sleepiness, driving, and motor vehicle crashes. Council on Scientific Affairs, American Medical Association. JAMA 1998;279(23):1908–13.
[4] Melamed S, Oksenberg A. Excessive daytime sleepiness and risk of occupational injuries in non-shift daytime workers. Sleep 2002;25(3):315–22.
[5] Carskadon MA, Dement WC, Mitler MM, Roth T, Westbrook PR, Keenan S. Guidelines for the multiple sleep latency test (MSLT): a standard measure of sleepiness. Sleep 1986;9(4):519–24.
[6] Scottish Intercollegiate Guidelines Network. Management of obstructive sleep apnoea/hypopnoea syndrome in adults. A national clinical guideline. Edinburgh: Scottish Intercollegiate Guidelines Network; 2003.
[7] Littner M, Johnson SF, McCall WV, Anderson WM, Davila D, Hartse SK, et al. Practice parameters for the treatment of narcolepsy: an update for 2000. Sleep 2001;24(4):451–66.
[8] Chesson Jr A, Hartse K, Anderson WM, Davila D, Johnson S, Littner M, et al. Practice parameters for the evaluation of chronic insomnia. An American academy of sleep medicine report. Standards of practice committee of the American academy of sleep medicine. Sleep 2000;23(2):237–41.
[9] Johns M, Hocking B. Daytime sleepiness and sleep habits of Australian workers. Sleep 1997;20(10):844–9.
[10] Liu X, Uchiyama M, Kim K, Okawa M, Shibui K, Kudo Y, et al. Sleep loss and daytime sleepiness in the general adult population of Japan. Psychiatry Res 2000;93(1):1–11.
[11] Johns MW. A new method for measuring daytime sleepiness: the Epworth sleepiness scale. Sleep 1991;14(6):540–5.
[12] Johns MW. Reliability and factor analysis of the Epworth sleepiness scale. Sleep 1992;15(4):376–81.
[13] Akashiba T, Tatsumi K, Chin K, Kimura H, Nishimura M, Hida W, et al. Current situation of the SAS diagnosis and treatment in facilities recognized by the Japanese respiratory society: results of the questionnaire survey. Nihon Kokyuki Gakkai Zasshi 2004;42(6):568–70 (in Japanese).
[14] Chin K, Fukuhara S, Takahashi K, Sumi K, Nakamura T, Matsumoto H, et al. Response shift in perception of sleepiness in obstructive sleep apnea–hypopnea syndrome before and after treatment with nasal CPAP. Sleep 2004;27(3):490–3.
[15] Takegami M, Sokejima S, Yamazaki S, Nakayama T, Fukuhara S. An estimation of the prevalence of excessive daytime sleepiness based on age and sex distribution of Epworth sleepiness scale scores: a population based survey. Nippon Koshu Eisei Zasshi 2005;52(2):137–45 (in Japanese).
[16] Adams R, Rosier M, Campbell D, Ruffin R. Assessment of an asthma quality of life scale using item response theory. Respirology 2005;10(5):587–93.
[17] Bjorner JB, Petersen MA, Groenvold M, Aaronson N, Ahlner-Elmqvist M, Arraras JI, et al. Use of item response theory to develop a shortened version of the EORTC QLQ-C30 emotional functioning scale. Qual Life Res 2004;13(10):1683–97.

[18] Revicki DA, Cella DF. Health status assessment for the twenty-first century: item response theory, item banking and computer adaptive testing. Qual Life Res 1997;6(6):595–600.
[19] Wunderlich GR, Evans KR, Sills T, Pollentier S, Reess J, Allen RP, et al. An item response analysis of the international restless legs syndrome study group rating scale for restless legs syndrome. Sleep Med 2005;6(2):131–9.
[20] Hambleton RK. Principles and selected applications of item response theory. In: Linn RL, editor. Educational measurement. New York: Macmillan; 1989. p. 143–200.
[21] van der Linden WJ, Hambleton RK, editors. Handbook of modern item response theory. Berlin: Springer; 1997.
[22] Muraki E. A generalized partial credit model: application of an EM algorithm. Appl Psychol Meas 1992;16:159–76.
[23] Acquadro C, Jambon B, Ellis D, Marquis P. Language and translation issues. In: Spilker B, editor. Quality of life and pharmacoeconomics in clinical trials. Philadelphia: Lippincott-Raven; 1996. p. 575–85.
[24] Flaherty JA, Gaviria FM, Pathak D, Mitchell T, Wintrob R, Richman JA, et al. Developing instruments for cross-cultural psychiatric research. J Nerv Ment Dis 1988;176(5):257–63.
[25] Bentler PM, Yuan KH. Test of linear trend in eigenvalues of a covariance matrix with application to data analysis. Br J Math Stat Psychol 1996;49(Pt. 2):299–312.
[26] Bock RD, Aitkin M. Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. Psychometrika 1981;46(4):443–59.
[27] Holland PW, Wainer H, editors. Differential item functioning. Hillsdale, NJ: Lawrence Erlbaum Associates; 1993.
[28] Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika 1951;16(3):297–334.
[29] Kramer MS, Feinstein AR. Clinical biostatistics. LIV. The biostatistics of concordance. Clin Pharmacol Ther 1981;29(1):111–23.
[30] Buysse DJ, Reynolds 3rd CF, Monk TH, Berman SR, Kupfer DJ. The Pittsburgh sleep quality index: a new instrument for psychiatric practice and research. Psychiatry Res 1989;28(2):193–213.
[31] Backhaus J, Junghanns K, Broocks A, Riemann D, Hohagen F. Test–retest reliability and validity of the Pittsburgh sleep quality index in primary insomnia. J Psychosom Res 2002;53(3):737–40.
[32] Doi Y, Minowa M, Uchiyama M, Okawa M. Development of the Japanese version of the Pittsburgh sleep quality index. Seishinka Chiryogaku 1998;13(6):755–63 (in Japanese).
[33] Doi Y, Minowa M, Uchiyama M, Okawa M, Kim K, Shibui K, et al. Psychometric assessment of subjective sleep quality using the Japanese version of the Pittsburgh sleep quality index (PSQI-J) in psychiatric disordered and control subjects. Psychiatry Res 2000;97(2–3):165–72.
[34] Johns MW. Sleep propensity varies with behaviour and the situation in which it is measured: the concept of somnificity. J Sleep Res 2002;11(1):61–7.
[35] Hagell P, Broman JE. Measurement properties and hierarchical item structure of the Epworth sleepiness scale in Parkinson's disease. J Sleep Res 2007;16(1):102–9.
[36] Johns MW. Sensitivity and specificity of the multiple sleep latency test (MSLT), the maintenance of wakefulness test and the Epworth sleepiness scale: failure of the MSLT as a gold standard. J Sleep Res 2000;9(1):5–11.