Please don't change the part in pink.
Please edit the title of no.18 of reference(p.15).

Article title:

Development of a Japanese version of ESS (JESS) based on Item Response Theory

Authors:

Misa Takegami, RN, MPH[1,2], Yoshimi Suzukamo, PhD[2,3], Hiroyuki Noguchi, PhD[4], Takafumi Wakita, MA[2,4], Kazuo Chin, MD, PhD[5], Hiroshi Kadotani, MD, PhD[6], Yuichi Inoue, MD, PhD[7], Takaya Nakamura, MD, PhD[8], Murray W. Johns, MD, PhD[9], Shunichi Fukuhara, MD, MSc[1,2]

Affiliations:

1. Department of Epidemiology and Healthcare Research, Graduate School of Medicine and Public Health, Kyoto University, Kyoto, Japan.
2. Department of Outcome Research, Institute for Health Outcomes and Process Evaluation Research, Kyoto, Japan.
3. Department of Physical Medicine and Rehabilitation, Graduate School of Medicine, Tohoku University, Sendai, Japan.
4. Department of Psychology and Human Development Sciences, Graduate School of Education and Human Development, Nagoya University, Aichi, Japan.
5. Department of Physical Therapeutics, Kyoto University Hospital of Medicine, Kyoto, Japan.
6. Horizontal Medical Research Organization, Graduate School of Medicine, Kyoto University, Kyoto, Japan.
7. Japan Somnology Center, Neuropsychiatric Research Institute, Tokyo, Japan.
8. Department of Respiratory Medicine, Kyoto University Graduate School of Medicine, Kyoto University Hospital, Kyoto, Japan.
9. Sleep Diagnostics Pty Ltd, Melbourne, Australia.

Disclosure statement:

This study was supported by grants from Institute for Health Outcomes and Process Evaluation Research (iHope International). Shunichi Fukuhara, Misa Takegami, Yoshimi Suzukamo and Takafumi Wakita have indicated no financial conflict of interest.

Correspondence and reprint requests to:

Misa Takegami

Graduate School of Medicine and Public Health, Kyoto University

Department of Epidemiology and Health Care Research

Yoshida Konoe-cho, Sakyo-ku, Kyoto 606-8501 JAPAN

Phone: +81-75-753-4645/Fax: +81-75-753-4644

E-mail address: takegami@pbh.med.kyoto-u.ac.jp

ABSTRACT

Objectives: In Japan, several localized version of the ESS are available. However, none of these were validated. This study aims to adapt the Epworth Sleepiness Scale (ESS) to Japanese culture, and to clarify the problems of unauthorized Japanese version of ESS. We then replaced problem items using Item Response Theory (IRT), and validated the Japanese version of ESS (JESS) using Classical Test Theory (CTT).

Methods: Based on discussion with the original developer, we translated the ESS items into Japanese and created an item pool for item replacement. Subjects (N=540) consisted of 85 patients with Obstructive Sleep Apnea Syndrome (OSAS), 54 patients with narcolepsy, 4 patients with idiopathic hypersomnia, and 397 healthy people. IRT was used to examine item characteristics for each item and to replace original items with adapted items in item pool. CTT was conducted to evaluate the psychometric characteristics of the JESS.

Results: During the translation phase of the ESS and pilot study, we identified two problematic items, partly because they were considered inappropriate to Japanese lifestyle. Using IRT, these two items were replaced with new items that had similar item characteristics to the original ones. Results for internal consistency (Cronbach's α=0.84) and test-retest reliability (ICC=0.75) confirmed the reliability of JESS, and factor analysis approved its construct validity. Responsiveness was confirmed by improvement of JESS scores after treatment in 32 OSAS patients.

Conclusion: Using IRT methodology, we succeed in replacing two problematic items of the JESS with new items without sacrificing scale equivalence to the original ESS.

249 words **(250 words)**

Key Words: the Epworth Sleepiness Scale (ESS), daytime sleepiness, cultural adaptation, Item Response Theory (IRT), validation study

INTRODUCTION

Daytime sleepiness is one of the most critical symptoms of sleep disorders. Since it can disrupt the patient's social life and poses a threat to traffic safety and public health, it is a serious problem not only for patients themselves but also for society as a whole[1-4]. Daytime sleepiness is also an important marker in the assessment of sleep disorders. Daytime sleepiness is measured subjectively as well as objectively. An objective measurement, the Multiple Sleep Latency Test (MSLT), is considered the gold standard[5] but has expensive cost and time requirements. The Epworth Sleepiness Scale (ESS) is a self-reported, subjective measurement[6,7]. The ESS consists of 8 items to measure sleepiness under 8 situations, with each item rated on 4-point scale (score 0–3 respectively) which are summed to give an overall score of 0-24. The normal score range is 2-10, with a higher ESS score indicating stronger subjective daytime sleepiness.

The ESS is the best available clinical tool to reveal the patient's perception of sleepiness, and is recommended in guidelines established to control sleepiness conditions such as Obstructive Sleep Apnea Syndrome (OSAS), narcolepsy, and insomnia[8-10]. The ESS has also been used in occupational study and community-based study[11,12].

In Japan, several localized version of the ESS are available. However, none of these were developed in accordance with standard international procedures or have been constructed content validity and are thus considered problematic. An advisory committee on sleep apnea syndrome within the Japanese Respiratory Society reported in 2004 that 167 of 277 hospitals used various versions of the translated ESS and in different ways (self-administration or interview). This inconsistency hampers comparison of patient sleepiness among hospitals[13]. In addition, although many papers that used ESS were published from Japan including us[14], there was a possibility that the measured conception was different with

localized ESS used in Japan and original ESS. Also, each item of the original ESS measures sleepiness under a daily life situation which depends on culture or lifestyle. This culturally dependent unfamiliarity has puzzled Japanese respondents and often leads to high rates of missing answers[15]. In particular, many patients cannot answer item 8 (In a car, while stopped for a few minutes in the traffic), and in fact some hospitals in Japan even omit item 8 in their assessment of sleepiness. Here, at the request of the Japanese Respiratory Society, we report the development of a Japanese version of the ESS (JESS) which was conducted with the collaboration of the original developer of the ESS, Johns, M. W.

Recently, Item Response Theory (IRT) has been increasingly used in studies of subjective scale construction, particularly in Quality of Life research. Common applications of IRT are the construction of short version scales, establishment of scoring algorithms, and Computerized Adaptive Testing (CAT) [16-18]. IRT describes the probability of an item response as a function of the subject's theta (latent trait) and item parameter (e.g. difficulty, discrimination). This provides a more precise examination of item characteristics than Classical Test Theory (CTT). Test information function identifies the range of the theta, allowing measurement accuracy to be evaluated at subjects' latent trait level. We utilized these characteristics of IRT to select alternative items from the item pool which have similar item characteristics to the original and are easier for Japanese to answer.

The purpose of this study was 1) to translate ESS into Japanese by means of the internationally adopted standard methods of scale development, and to clarify the problems of unauthorized Japanese version of ESS, 2) to apply Item Response Theory (IRT) to select alternative items equivalent to the original ESS items but more appropriate to Japanese lifestyle, and 3) to confirm the psychometric properties of the JESS such as reliability and validity based on Classical Test Theory (CTT).

METHODS

Translation and preparation of item pool

The Japanese version was developed in conformance with standard methods that have been adopted internationally[19]. The process includes forward translation, back translation, examination of translation quality and a pilot test on 10 persons.

First, at the forward step, two types of equivalence were examined: 1) conceptual equivalence, which determines that the instrument is measuring the same theoretical construct in each culture; and 2) technical equivalence, which determines that the methods of assessment are comparable in each culture with respect to data collection[20, 21]. Second, two bilingual translators - Japanese native speakers with a good command of English - translated 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 keeping semantic equivalence[20, 21]. Content equivalence, which determines that the content of each item is relevant to the phenomenon of each culture, was also examined. Additionally, new items which better fit Japanese culture and lifestyle were added to replace problematic items that frequently had missing values in a community-based survey[15]. The pilot study confirmed the appropriateness of the new items, which were selected based on a similar difficulty to replaced original item of the possibility of dozing off.

Back translation was conducted by a different bilingual translator, who was a native English speaker. We referred this back-translated version to the original developer of the ESS and discussed differences between the translated and original versions as well as lifestyle differences between English-speaking countries and Japan. We kept repeating arguments about translational problems which were pointed out by the original developer. We took the process of translation from forward step until approved by the developer. This

process was continued until the authorized JESS was finalized.

Subjects and data collection

Cross-sectional data were obtained from a convenience sample of 143 outpatients and 397 healthy people, who completed a self-administrated questionnaire between October 2004 and December 2005. The outpatients had one of following sleep disorders: OSAS, narcolepsy, or idiopathic hypersomnia. They were either at their first visit or under treatment at either of two outpatient hospitals specializing in sleep disorders. Continuous Positive Airway Pressure (CPAP) is the most commonly used treatment for OSAS and shows improved outcomes over other treatments. We therefore restricted OSAS patients to those receiving CPAP treatment. The healthy subjects group comprised male employees, undergraduate students, and retirees.

Measured items were the translated JESS, sex, date of birth, medical history, height, weight, sleep duration, the absence or presence of snoring, and whether the subject possessed a driver's license. For the outpatient group, clinicians provided the degree of severity before treatment, treatment method and period, blood pressure, and physiological indexes related to sleep disorders, including apnea-hypopnea index, oxygen desaturation index, and minimum SaO_2

To examine test-retest reliability, the male employees in the sample completed a questionnaire which included the translated ESS including nee items only one week after the first data collection. Additionally, to examine responsiveness, 32 OSAS patients, undergoing CPAP treatment, were subject to one month follow-up research.

This study was authorized by the Institutional Review Board of Kyoto University Graduate School of

Medicine and the Institutional Review Board of the Neuropsychiatric Research Institute.

Data analysis

Data analysis consisted of two steps: 1) selection of items for replacement based on IRT, and 2) evaluation of the psychometric properties of the JESS based on CTT. We used Parscale 4.1 (SSI, Inc., 2003) and SPSS12.0J for Windows (SPSS, Inc., 2003) for data analysis.

Item selection for equivalent item replacement using IRT

A precondition of the IRT analysis is the uni-dimensionality of the item pool. This was examined by a scree test[22] of eigenvalues in factor analysis. The Generalized Partial Credit Model (GPCM)[23] was used to analyze ordered response categories by the IRT model. The GPCM estimates three kinds of parameters of item characteristics, namely the slope parameter, the location parameter, and the category parameter. The slope parameter describes the discriminating power of the item: items with higher slope parameters are better at discriminating the strength of drowsiness than those with a lower slope parameter. The location parameter is conceptually similar to item difficulty. The category parameter corresponds to each response category and describes item difficulty in conjunction with the location parameter[24, 25]. Item characteristics confirmed equivalent if these parameters for the new items are similar to those of the originals[24, 25]. Category parameter is assumed to be the same for all items with the same response categories in GPCM. The item parameters were estimated using marginal maximum likelihood estimation[26]. Item selection was conducted after in-depth examination of item characteristics in the item pool based on the location and slope parameters.

Finally, the selected items were confirmed by the Item Response Category Characteristic Curve (IRCCC). The IRCCC describes the response probability for each item category as a function of slope

parameter, location, and theta. Theta describes the value of the latent trait, with a 0 theta meaning average ability for the target group. IRT advances the concept of test information as a replacement for test reliability. Test information is a function of the model parameters[24, 25], which define the Standard Error (SE) of measurement for each ability level. The test information curve of the JESS was therefore compared to that of the preliminary JESS before item selection.

Psychometric properties of the new JESS using CTT

Item analysis

Items with a missing value rate of 10% or more were examined. Further, mean and standard deviation of the JESS total score was examined. Response rates for each item were calculated to identify items with a response category with a 90% or greater response rate.

Reliability

Cronbach's alpha coefficient[27] was used to examine if the internal consistency reliability of the JESS was 0.7 or higher. Additionally, test-retest reliability was examined if Intra-class Correlation Coefficient (ICC) [28] of the JESS with a one-week interval was 0.7 or higher.

Validity

Factor analysis was used to confirm the uni-dimensionality, which indicates factor validity. Two types of criterion-based validity were examined. First, concurrent validity was examined using the association between daytime dysfunction domain with the Pittsburg Sleep Quality Index (PSQI) [29-31] and JESS. Daytime dysfunction measures dysfunction in daytime activities such as difficulty in staying awake during social activities and lack of sufficient motivation to get things done. The domain of PSQI is by a 4-point scale (score 0-3), with a higher score indicating stronger daytime dysfunction. Second, known-groups validity was examined by comparisons between patients and healthy people, patients before and under treatment, and patients using and not using sleeping pills. We used analysis of covariance to estimate unadjusted mean difference among each group. We estimated least-squares means

of JESS score with adjustment of age and sex and used Bonferroni's test to control for multiple comparisons.

Responsiveness

Responsiveness was examined using data from OSAS patients who completed the survey at baseline and at 1 month after CPAP treatment. Differences related to CPAP in these patients were analyzed with Student's t-test for paired data.

RESULTS

Translation and item pool

The most important change during the translation procedure was made on the instruction and response sentences. Although the original ESS describes sleepiness into "doze off or fall asleep", "doze off" are very different from "fall asleep" in Japanese. The original author pointed out that "dozing off" meant sleepiness for a several or more seconds. We therefore defined this as "doze off (sleep for several seconds or several minutes)". Importantly, the author pointed out that the ESS was intent to ask not "how often" but "how likely" they were to doze off in each situation, although several localized versions of the ESS have asked frequency of falling asleep in Japan. We decided that the question sentence and responses need to phrased in terms of the possibility rather than the fact, so we changed these to "How much is the possibility of your dozing off?", "There is no possibility of dozing off" and so on. We also added postscripts that "It is very important that you answer all items. Please try, as much as possible, to answer all of the items." Those changes seem to have contributed to a decrease in the response failure rate (see results of item analysis).

In-depth discussion was conducted with the original developer to confirm the situation and wording of each item, including new items. According to the original developer, item 8 describes the situation of being in a car not only as the driver, but also as a passenger. In the JESS, however, item 8 (in a car) had been mistranslated to "while driving" or "when driving a car", which confined subjects in Japanese to drivers only, and discussion on this revealed a serious translation problem. In fact, the original English version has been criticized over the difficulty in precisely understanding the meaning of item 8, in particular its vagueness concerning driver vs. passengers. Consequently, we decided to replace item 8 with one involving a different situation and activity. Six items were authorized for addition to the item pool as candidate replacements for the original item 8. We selected situations such as dozing while

eating, soaking in the bathtub, and handwriting as candidates for an alternative, because they may cause dozing off as same as the situation in a car.. Additionally, discussion with the original developer suggested that item 1, which was translated as "while sitting and reading a book", confined subjects to reading a book only and was associated with a relatively high response failure. Thus, item 9 was added as a candidate item to replace item 1.

Subject characteristics

Subject characteristics are shown in Table 1 (N=540). The healthy subject groups (n=397; 75.9% male, average age 39.0±15.0) consisted of 229 male salaried employees, 125 undergraduate students, and 43 retirees. The patients groups (n=143; 84.7% male; average age 33.8±15.7) consisted of 54 patients with OSAS, 54 with narcolepsy, and 4 with idiopathic hypersomnia.

Item selection using IRT

The scree plots showed that attenuation of eigenvalues for factor analysis were 6.19, 1.51, 1.09, 0.97, 0.70, which indicated a one-factor solution. At the same time, all factor loadings for the items met the criterion (0.50~0.77). These results confirmed the uni-dimensionality of the JESS.

Table 2 shows the slope parameter and location parameter as estimated by GPCM. According to these, item 1, "While sitting and reading a book" (slope=0.81, location=0.36), was replaced with item 9, "While reading something in a chair (newspapers, magazines, books, documents, etc)" (slope=1.16, location=0.66). Also, item 8, "In a car, while stopping for a few minutes because of a signal or a traffic jam" (slope=0.57, location=2.27) was replaced item 13, "While sitting and handwriting" (slope=0.75, location=2.00). Comparisons between before and after items are shown in Figure 1. Items 1 and 9, as well as items 8 and 13, showed a relatively similar IRCCC shape. Additionally, the use of items with a

high slope parameter increased the discriminating power of the response alternatives, which confirmed the accuracy of the scale.

As shown in Figure 2, comparison of the test information curve between the original version with original items and the JESS replaced 2 items confirmed that the amount of test information provided by the JESS was higher than that of the original version. Especially, test information is higher at the level of strong sleepiness. For instance, amount of test information of the revised JESS 1.45 higher than that of the original version (amount of test information=7.79, 6.34, respectively).

Psychometric properties of new version JESS using CTT

Item analysis

The range of missing value rates for the JESS was 0.9-2.2% (1.3% for replaced item 1 and 1.5% for replaced item 8). Results showed that the item 8 reduced the missing values rate by item replacement (5.2% for original item 8 vs. 1.5% for replaced item 8). No item showed a 90% or greater response in one response category. The mean JESS total score was 9.0 (SD= 4.7).

Reliability

Cronbach's alpha coefficient for the JESS total score was 0.84. When the alpha coefficient was calculated for each of the 8 items by eliminating each item, one by one, the range was 0.81-0.83, and no items were found to change the internal consistency substantially. Test-retest reliability was ICC=0.75 (n=53). These results showed that the JESS had high reliability.

Validity

Results of factor analysis showed that item factor loadings were 0.56~0.74. Further, individual factor analyses for outpatient and healthy subject groups confirmed the uni-dimensionality of JESS in each groups. The proportion of factor contribution was 47.8% and item-total correlation coefficients were

0.58~0.76. The JESS total score was 7.0 for the group with the best daytime function and 13.7 for that with the worst daytime function. This difference indicated that the JESS total score increased significantly as daytime dysfunction worsened. Figure 3 shows the dose-response relation between JESS and daytime dysfunction by the PSQI (ANCOVA p-value for trend, $p<0.001$), which confirmed the concurrent validity of the JESS. On comparison among the healthy subject and OSAS and narcolepsy patient groups to examine known-groups validity showed that both the OSAS and narcolepsy patients had significantly stronger drowsiness than healthy people ($p=0.044$, $p<0.001$. respectively) (Figure 4). For comparisons within each patient group, OSAS patients before treatment (9.2) showed significantly stronger drowsiness than OSAS patients under treatment (6.6), and OSAS patients not taking sleeping pills (9.0) showed significantly stronger drowsiness than those taking sleeping pills (11.1) ($p<0.024$, $p<0.003$, respectively).

Responsiveness

OSAS patients (n=27) showed a significant change in JESS total scores before and after treatment ($p<0.001$; before treatment, mean=10.0 and SD=0.42; after treatment, mean=6.0 and SD=4.19).

DISCUSSION

In this study, it became clear that the unauthorized Japanese translations of the ESS used in Japan measured different constructs than the original ESS. Moreover, we found that the use of item 8 (In a car) was difficult to answer and its replacement was better in Japan.

On discussion with the author who developed the original ESS during our translation process, these Japanese versions posed questions in the factual terms of "how often," and measured more severe sleepiness. In contrast, the original ESS asked questions in terms of probability, i.e. "how likely". Lack of content equivalence is the most serious problem confronting the development of foreign scales. To overcome this problem, we were here assured of the importance of the process to examine content validity, including forward translation, back translation, examination of translation quality and a pilot test, in translation and cultural adaptation.

Using the IRT model, we replaced item 8 in the original ESS (In a car), which had been the most problematic in clinical settings, with an equivalent item asking about sleepiness in a situation familiar to Japanese. IRT results indicated that the original ESS included items with various item difficulties. The new items were selected to maintain the same rank order by location parameter. Examination of means for each showed that item 6 (talking with someone) and replacement item 8 (sitting and handwriting) maintained their original rank order of first and second, respectively. However, item 2 (watching TV) and replacement item 1 (reading something) exchanged their order of third and fourth. This suggested that the original item 1 was handicapped by its apparent exclusion of reading materials other than books. Although replacement item 8 was markedly different to the original, it had similar difficulty but improved discrimination, demonstrating the benefit of this item replacement. Compared to the original version before item replacement, the JESS after it showed larger amount of test information. Further, it

also shows higher test information at the level of strong sleepiness, suggesting that the JESS will be useful for patients with strong drowsiness. These results confirm the advantage of the JESS over previous unauthorized translations.

On comparison of IRT results between people with and without a driver's license, location parameter for the original item 8 was 1.99 (SE=0.12) for people with a license and 3.67 (SE=1.04) for those without. This has been referred to as differential item functioning (DIF) [32], which refers to the unexpected difference in item performances among groups of equally sleepiness, usually classified by ethnicity or gender. The existence of DIF justified the necessity of replacing the original item 8 with an equivalent item whose response was independent from subject attributes. The present study used a convenience sample resulted in high response rate for each JESS item. However, a localized translated Japanese version of the ESS showed high missing rate in community-based surveys[15]. Further research is needed to examine whether these item replacements in the JESS are effective in reducing response failure.

The psychometric properties of JESS were examined by CCT. The results supported its internal consistency reliability, test-retest reliability, factor validity, and responsiveness. No significant differences were seen between treated and untreated narcolepsy patients. The reason why criterion-based validity was not statistically confirmed might be the small sample size or the smaller treatment effect of narcolepsy patients than OSAS patients. Nevertheless, we confirmed sufficient criterion-based validity for the other hypotheses.

The strength of this study is its use of the IRT model for item selection. IRT has advantages in describing item characteristics in more depth than CCT and in comparing each item by means of location parameter, which means difficulty, on a common scale, theta[24, 25]. In the case of CCT, item

difficulty is defined by the average of the response choice's value (0-3). However, the interval of the response choices is not necessarily equal. In the context of CCT, this means that the comparison of item category parameters does not mean comparison of item difficulties. In addition, IRT has advantages in examining whether the precision of the JESS improved after item replacement using test information. These characteristics suggest that IRT is a better option for determining item replacement than CCT. Moreover, to improve risk evaluation of sleepiness with ESS, additional steps to select items preventing response bias using CAT advanced IRT should be examined. The correlation between sleepiness and traffic accidents has become an important public safety issue[3,4]. In 2002, the Japanese traffic laws were revised to add "sleep disorders with serious symptoms of drowsiness" to those disorders for which a driving license can be refused or disqualified. ESS is used to evaluate the risk of sleepiness, but the fact that it is self-administered raises strong concern over bias. So, the further study is required using IRT.

Scale translations developed in foreign countries require five types of equivalence for cultural adaptation[21]. One of them, content equivalence, requires investigation to determine if the translated scale is culturally appropriate while maintaining equivalent meaning in two cultures. Since each item of the ESS asks about a daily life situation which depends strongly on culture, the use of different translation vocabularies is not sufficient to ensure equivalence. We therefore used IRT to select new items to replace those inappropriate to Japanese culture. The present results confirmed that the item characteristic function of the replaced items were equivalent to those of the original. Nevertheless, although there is no warranty that the same result for these replacements would be obtained in other cultures or countries, replacement with familiar items has practically advantages in reducing the participant's burden and in improving the accuracy of measurement compared with the case where the item is not replaced.

This study had three limitations. First, it did not include all sleep disorders that cause daytime sleepiness

(e.g. insomnia, periodic limb movement disorder and other miscellaneous disorders). Therefore, whether these findings are applicable to patients with disorders other than OSAS and narcolepsy remains to be studied. Second, only daytime sleepiness was measured, so no discussion can be made about the association between MSLT and JESS. Differences between the evaluation of sleepiness in daily life and in dark laboratory settings have been suggested [33]. The association with MSLT is a contentious issue, so further research is needed. Finally, this study used a convenience sample, and thus might not reflect results with a general population. Confirmation of these results will require further research in other sleep disorder patients as well as in general populations.

The ESS has been translated into many languages and used as common measurement of daytime sleepiness in not only English-speaking but also many other countries[34-38]. It is impossible to directly compare ESS raw scores among different countries. However, for countries with similar SD values, norm-based scoring on the basis of each country's general population data is feasible. This is the same method used for the 36-item survey Short-Form Health Survey to measure Health-related Quality of life[39]. Norm-based scoring is also useful in minimizing "cultural response bias," which means the average responses may differ between groups if people of different countries or cultures are asked to judge, for example, sleepiness under the same circumstances. However, cultural adaptation of scales has to be confirmed before norm-based scoring can be calculated. For the Japanese translation, the item measuring sleepiness while driving was problematic and required replacement for cultural adaptation. Given the possibility that other countries or cultures have similar problems, re-examination of the suitability of each item may be required.

CONCLUSION

We completed a standardized translation, adaptation and validation of the ESS in Japan. To our

knowledge, this study was the first application of IRT to culturally relevant item selection for the ESS. Following replacement with alternative items, the JESS is now characterized by content equivalence and adaptation to Japanese culture, and is thus expected to solve the problems of scale application, such as in reducing response failure. This study demonstrated the potential of IRT in the selection of items to solve this type of cultural adaptation problem.

ACKNOWLEGEMENTS

We are grateful to Yasunori Oka, Itsunari Minami and Yuriko Nakayama for recruiting subjects and collecting data. We also wish to thank Tsutomu Namikawa for suggestion on this analysis.

REFERENCES

1. Akashiba T, Kawahara S, Akahoshi T, et al. Relationship between quality of life and mood or depression in patients with severe obstructive sleep apnea syndrome. Chest 2002;122:861-5.

2. Melamed S, Oksenberg A. Excessive daytime sleepiness and risk of occupational injuries in non-shift daytime workers. Sleep 2002;25:315-22.

3. Lyznicki JM, Doege TC, Davis RM, et al. Sleepiness, driving and motor vehicle crashes. JAMA 1998;279:1908-13.

4. Findley LJ, Fabrizio M, Thommi G, et al. Severity of sleep apnea and automobile crashes. N Engl J Med 1989;320:868-9.

5. Carskadon MA. Dement WC. Mitler MM, et al. Guidelines for the multiple sleep latency test (MSLT): a standard measure of sleepiness. Sleep 1986;9:519-24.

6. Johns MW. A new method for measuring daytime sleepiness: the Epworth Sleepiness. Sleep 1991;14:540-545.

7. Johns MW. Reliability and Factor Analysis of the Epworth Sleepiness Scale. Sleep 1992;15:376-381.

8. Scottish Intercollegiate Guidelines Network. Management of obstructive sleep apnoea/hypopnoea syndrome in adults. A national clinical guideline. Edinburgh: Scottish Intercollegiate Guidelines Network, 2003.

9. Littner M, Johnson SF, McCall WV, Anderson WM, Davila D, Hartse SK, Kushida CA, Wise MS, Hirshkowitz M, Woodson BT. Practice parameters for the treatment of narcolepsy: an update for 2000. Sleep 2001;24:451-66.

10. Chesson A Jr, Hartse K, Anderson WM, Davila D, Johnson S, Littner M, Wise M, Rafecas J. 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;23:237-41.

11. Johns MW, Hocking B. Daytime sleepiness and sleep habits of Australian workers. Sleep 1997;20:844-9.

12. Liu X, Uchiyama M, Kim K, et al. Sleep loss and daytime sleepiness in the general adult population of Japan. Psychiatry Res 2000;93:1-11.

13. Akashiba T, Tatsumi K, Chin K, Kimura H, Nishimura S, Hida W, Fukuhara S, Fujimoto K, Mishima M, Horie T. Advisory committee for sleep apnea syndrome. 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;42:568-570. (in Japanese)

14. Chin K, Fukuhara S, Takahashi K, Sumi K, Nakamura T, Matsumoto H, Niimi A, Hattori N, Mishima M, Nakamura T. 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;52:137-45. (in Japanese)

16. 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;6:595-600.

17. Adams R, Rosier M, Campbell D, et al. Assessment of an asthma quality of life scale using item-response theory. Respirology 2005;10:587-93.

18. Bjorner JB, Petersen MA, European Organisation for Research and Treatment of Cancer Quality of Life Group, et al. Use of item response theory to develop a shortened version of the EORTC QLQ-C30 emotional functioning scale. Qual Life Res 2004;13:1683-97.

19. Acquadro C, Jambon B, Ellis D, et al. Language and Translation Issues. In: Bert Spilker, eds. Quality of Life and Pharmacoeconomics in Clinical Trials, 2nd ed. Philadelphia: Lippincott-Raven Publishers, 1996:575-85.

20. Cramer JA, Spilker, B. Reported methodology. In: Quality of Life and Pharmacoeconomics: An Introduction. Lippincott-Raven Publishers, Philadelphia, 1997: 96-149.

21. Flaherty JA, Gaviria FM, Pathak D,et al. Developing instruments for cross-cultural psychiatric research.. J.Nerv Ment Dis 1988; 176: 257-263.

22. 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;49:299-312.

23. Muraki E. A generalized partial credit model. In: van der Linden WJ, Hambleton RK, eds. Handbook of Modern Item Response Theory. Berlin: Springer, 1997:153–64.

24. Hambleton RK. Principles and Selected Applications of Item Response Theory. In: Linn RL eds. Educational Measurement. New York: Macmillan, 1989:143–200.

25. van der Linden WJ, Hambleton RK. Handbook of Modern Item Response Theory. Berlin: Springer, 1997.

26. Bock RD, Aitkin M. Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika 1981;46:443–59.

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

28. Kramer MS, Feinstein AR. Clinical biostatistics LIV: the biostatistics of concordance. Clin Pharmacol Ther 1981;29:111-23.

29. Buysse DJ, Reynolds CF, Monk TH, et al. The Pittsburgh Sleep Quality Index: A New Instrument for Psychiatric Practice and Research. Psychiatry Res 1988;28:193-213.

30. 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;53:737-40.

31. Doi Y. Minowa M, Uchiyama M, 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;97:165-72.

32. Holland PW, Wainer H. Differential Item Functioning. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc., 1993.

33. John 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: 5-11.

34. KF Chung. Use of the Epworth Sleepiness Scale in Chinese patients with obstructive sleep apnea and normal hospital employees. J Psychosom Res 2002;49:367-72.

35. Hublin C, Kaprio J, Partinen M, et al. Daytime sleepiness in an adult Finnish population. J Intern Med 1996;239:417-23.

36. Souza JC, Magna LA, Reimao R. Excessive daytime sleepiness in Campo Grande general population, Brazil. Arq Neuropsiquiatr 2002;60:558-62.

37. Tsara V, Serasli E, Amfilochiou A, Constantinidis T, Christaki P. Greek version of the Epworth Sleepiness Scale. Sleep Breath 2004;8:91-5.

38. Gander PH, Marshall NS, Harris R, Reid P. The Epworth Sleepiness Scale: influence of age, ethnicity, and socioeconomic deprivation. Epworth Sleepiness scores of adults in New Zealand. Sleep 2005 1;28:249-53.

39. Aaronson NK, Acquadro C, Alonso J, et al: International Quality of Life Assessment (IQOLA) Project. Qual Life Res 1:349–51, 1992

Please don't change the part in pink.

Table 1. Subject characteristics

	n=540
Male, number (%)	395 (73.1%)
Age, mean (min – max)	41.4 (16 – 82)
Body mass index, mean±SD	23.5±4.2
Snore (yes, everyday or often)	209 (38.7%)
Sleep medication use (yes)	48 (0.1%)
Smoking (yes)	174 (32.2%)
Drinking (almost every day)	190 (35.2%)
Driver's license (yes)	438 (81.1%)
Disease	
OSAS (before treatment)	53 (9.8%)
OSAS (under treatment)	32 (5.9%)
narcolepsy (before treatment)	11 (2.0%)
narcolepsy (under treatment)	43 (8.0%)
idiopathic hypersomnia (before treatment)	4 (0.7%)

Table 2. Item characteristics estimated by IRT analysis

	a	SE	b	SE
Original Items				
1. While sitting and reading a book	0.81	0.06	0.36	0.05
2. While watching a television	0.74	0.06	0.69	0.05
3. While sitting at a meeting or a theater without actively speaking, etc	0.74	0.06	0.17	0.05
4. While being in a car as a passenger for one consecutive hour	0.73	0.06	-0.05	0.05
5. While lying down and taking a rest in the afternoon(s)	0.71	0.06	-0.94	0.06
6. While sitting and talking with someone	0.91	0.09	2.36	0.10
7. While sitting quietly after taking lunch (without liquors)	0.87	0.06	0.08	0.04
8. In a car while stopping for a few minutes because of a signal or a traffic jam	0.57	0.06	2.27	0.14
Additional items				
9. While reading something in a chair (newspapers, magazines, books, documents, etc)	1.16	0.08	0.66	0.04
10. While eating with your family or acquaintances	1.03	0.12	2.95	0.14
11. While sitting in a train, bus, or a taxi without talking to anyone	0.83	0.06	0.02	0.04
12. While soaking in the bathtub during bath	0.44	0.05	1.72	0.13
13. While sitting and handwriting	0.75	0.07	2.00	0.09
14. While eating alone, seated	0.84	0.10	2.90	0.15
15. While sitting and waiting for a train or a bus without talking to anyone	0.55	0.05	1.71	0.10

* a=slope parameter, b=location parameter, SE=standard error

* Category parameter: 1.23, -0.39, -0.84

Table 3. Factor loadings for each item in the revised JESS items for all subjects, the patient group and the healthy subjects group.

ESS item no.	All subjects (n=513)	Patient group (n=134)	healthy subjects group (n=379)
1	0.731	0.801	0.659
2	0.592	0.636	0.536
3	0.636	0.623	0.630
4	0.642	0.660	0.628
5	0.617	0.602	0.648
6	0.554	0.656	0.445
7	0.701	0.742	0.691
8	0.566	0.654	0.459

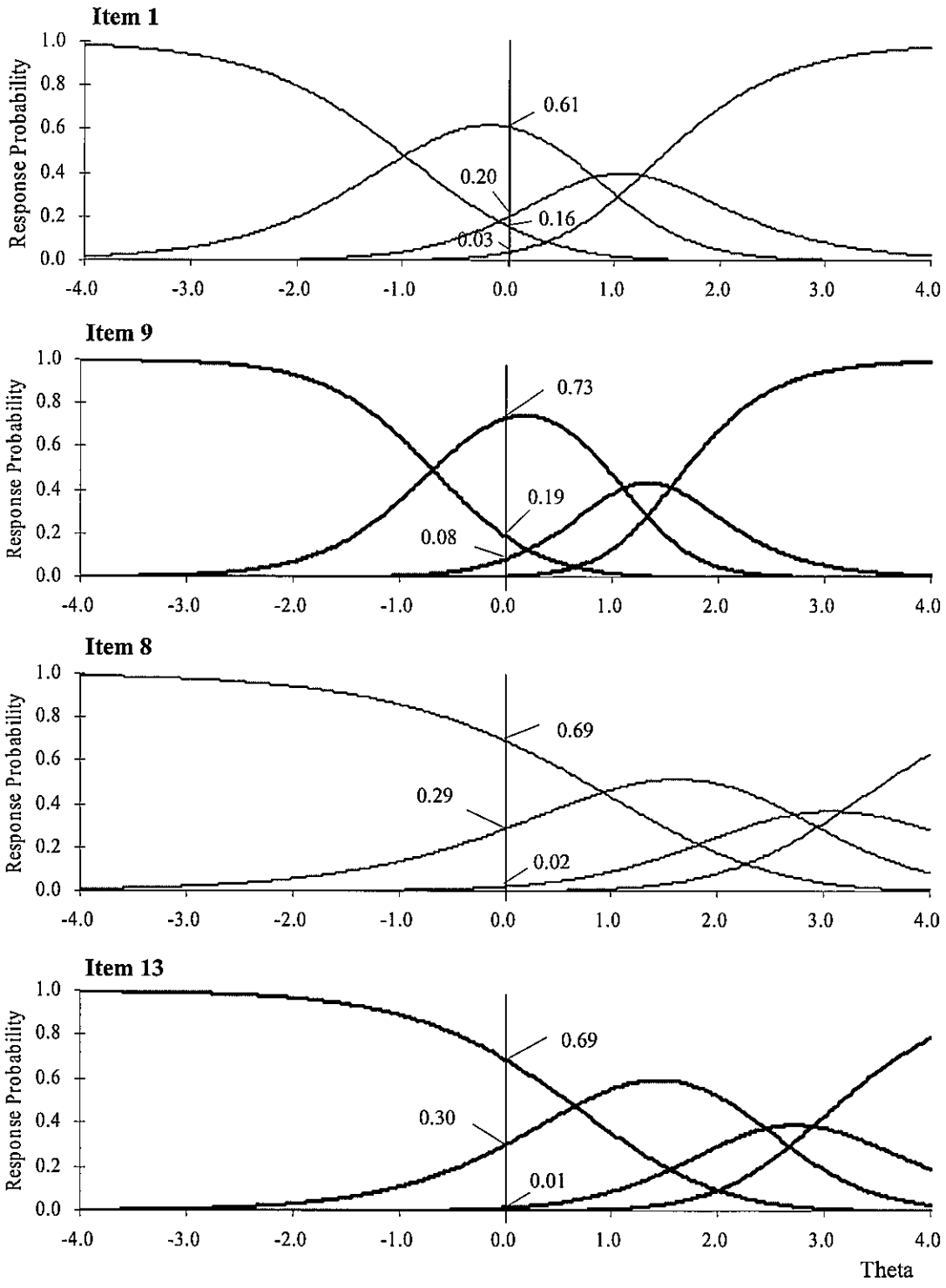

Figure 1. Item response category characteristic curve used by Generalized Parcial Credit Model. (Category parameter: 1.23, -0.39, -0.84) The curve lines represent response probability functions for each response choice. Item 1 was replaced with tem 9. Item 8 was replaced with item13.

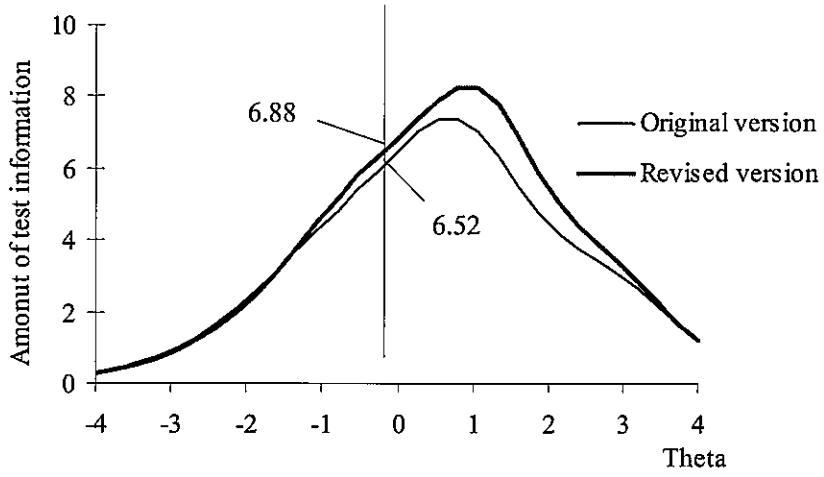

Figure 2. Comparison of test information curve between original version and revised version. Original version means the JESS with original items. Revised version means the JESS with two replaced items by results of IRT.

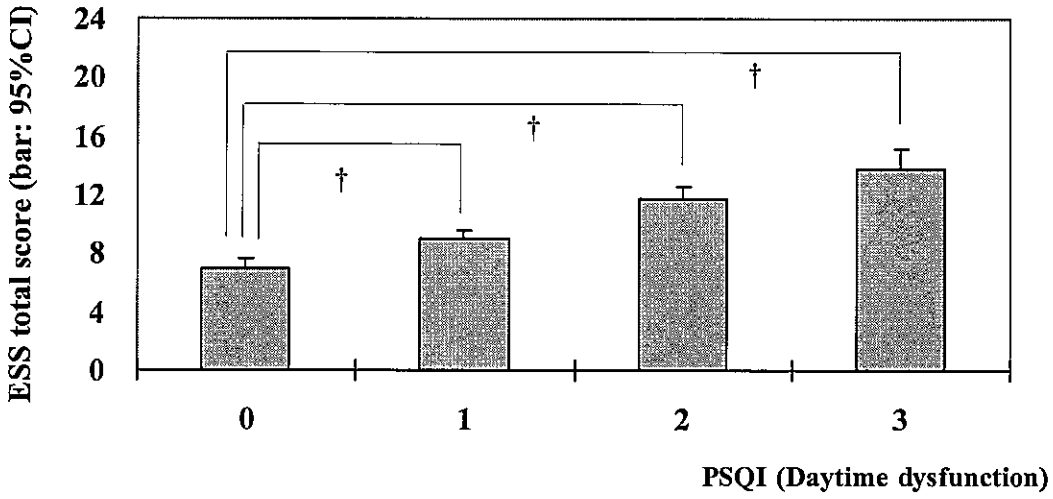

Figure 3. Associations between ESS total score and dysfunction domain of Pittsburg sleep quality index (PSQI) after adjustment for sex and age. Dysfunction domain score in PSQI range from 0 to 3, with a higher score indicating severer daytime dysfunction.
†: $p<0.001$, ANCOVA p-value for trend $p<0.001$

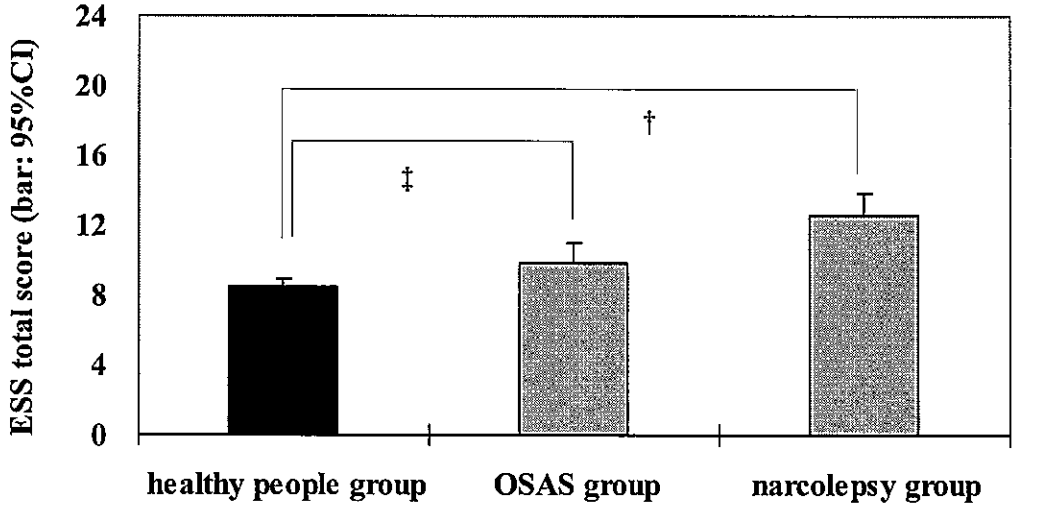

Figure 4. The adjusted mean ESS total scores and 95% confidence interval (95%CI) of healthy people group, obstructive sleep apnea (OSAS) group and narcolepsy group. ESS total score adjusted for sex and age.
†: $p<0.001$, ‡: $p=0.044$

Faxed 27.9.06.