Psychological Assessment
2003, Vol. 15, No. 2, 137–148

1040-3590/03/$12.00 DOI: 10.1037/1040-3590.15.2.137

Reliability and Validity of the Women's Health Initiative Insomnia Rating Scale

Douglas W. Levine
Wake Forest University School of Medicine

Daniel F. Kripke and Robert M. Kaplan
University of California, San Diego

Megan A. Lewis
University of North Carolina at Chapel Hill

Michelle J. Naughton
Wake Forest University School of Medicine

Deborah J. Bowen
Fred Hutchinson Cancer Research Center

Sally A. Shumaker
Wake Forest University School of Medicine

The reliability and construct validity of the 5-item Women's Health Initiative Insomnia Rating Scale (WHIIRS) were evaluated in 2 studies. In Study 1, using a sample of 66,269 postmenopausal women, validity of the WHIIRS was assessed by examining its relationship to other measures known to be related to sleep quality. Reliability of the WHIIRS was estimated using a resampling approach; the mean alpha coefficient was .78. Test–retest reliability coefficients were .96 for same-day administration and .66 after a year or more. Correlations of the WHIIRS with the other measures were in the predicted directions. Study 2 used a sample of 459 women and compared the WHIIRS with objective indicators of sleep quality. Results showed that differences in the objective indicators could be detected by the WHIIRS. Findings suggest that a between-group mean difference of approximately 0.50 of a standard deviation on the WHIIRS may be clinically meaningful.

In the early 1990s, the National Institutes of Health was planning and developing the Women's Health Initiative (WHI). This study is possibly the world's largest clinical investigation of the determinants of the common causes of morbidity and mortality in postmenopausal women 50–79 years of age. In all, 161,809 women were enrolled in the various arms of this 15-year study that ends in 2007. Because of the prevalence and importance of sleep disorders (e.g., Floyd, Medler, Ager, & Janisse, 2000; Foley, Monjan, Izmirlian, Hays, & Blazer, 1999; Ford & Kamerow, 1989; Maggi et al., 1998; Ohayon, Caulet, & Lemoine, 1998; Schwartz et al., 1999), and because epidemiologic studies often show that women and older persons are more likely to have sleep disorders and accompanying psychological distress, somatic anxiety, major depression, and multiple health problems, the WHI decided to include a measure of sleep quality (e.g., Ford & Cooper-Patrick, 2001; Mellinger, Balter, & Uhlenhuth, 1985; Ohayon, 2002; Sateia, 2002; Sateia, Doghramjii, Hauri, & Morin, 2000). At the time (i.e., the early 1990s), there was no widely used, short, reliable, and valid scale available.[1] Thus, the WHI Program Council, in consultation with sleep experts, chose to develop its own set of items to be used in the study.

Ten sleep-related questions were asked of participants. These items were intended to assess medication use or sleeping aids, somnolence or daytime sleepiness, napping, sleep initiation insomnia or sleep latency, sleep maintenance insomnia, early morning awakening, snoring (an indicator of sleep-disordered breathing), perceived adequacy of sleep or sleep quality, and sleep duration or quantity. Using these 10 items, Levine et al. (2003) developed a five-item scale that they called the Women's Health Initiative Insomnia Rating Scale (WHIIRS). The WHIIRS assesses insomnia symptoms (i.e., sleep latency, sleep maintenance, early morning awakening, and sleep quality), and items are shown in the Appendix.

With a sample of almost 70,000 WHI participants, Levine et al. (2003) used a novel resampling method to conduct an extensive investigation of the factor structure of the instrument. To develop the scale, 120,000 separate factor analyses were conducted with 1,000 women in each sample. The resulting scale was a measure of perceived insomnia symptoms. Content validity was

Douglas W. Levine, Michelle J. Naughton, and Sally A. Shumaker, Department of Public Health Sciences, Wake Forest University School of Medicine; Daniel F. Kripke, Department of Psychiatry, University of California, San Diego; Robert M. Kaplan, Department of Family and Preventive Medicine, University of California, San Diego; Megan A. Lewis, Health Behavior and Health Education, University of North Carolina at Chapel Hill; Deborah J. Bowen, Cancer Research Prevention, Fred Hutchinson Cancer Research Center, Seattle, Washington.

This work was supported by the National Institutes of Health (Women's Health Initiative, Grants HL55983, HL62180, and AG15763). We thank Joseph D. Assmus for his contributions to Study 2 and Ute Bayen for his helpful comments.

Correspondence concerning this article should be addressed to Douglas W. Levine, Section on Social Sciences and Health Policy, Department of Public Health Sciences, Wake Forest University School of Medicine, Winston-Salem, North Carolina 27157. E-mail: dlevine@wfubmc.edu

[1] The Pittsburgh Sleep Quality Index was then relatively new, was not in wide use, and had been validated on a relatively small sample.

assessed by comparing the WHIIRS with the definitions of insomnia included in the major nosologies (i.e., *International Classification of Sleep Disorders* [American Academy of Sleep Medicine, 1997], *Diagnostic and Statistical Manual of Mental Disorders* [4th ed.; American Psychiatric Association, 1994], and *International Classification of Diseases* [10th ed.; World Health Organization, 1992]). This assessment revealed that the WHIIRS items corresponded to the majority of insomnia characteristics noted in the nosologies and the literature. The factor analyses showed that the WHIIRS had desirable measurement properties. The scale has a stable one-factor solution, and multigroup structural equation modeling revealed measurement invariance across age and race–ethnic groups. Norms for this scale were provided by age and race–ethnic groups.

The purpose of this study was to evaluate the reliability and validity of the WHIIRS. Two studies based on data from WHI participants are reported. In the first study, a detailed analysis of the reliability (both internal consistency and test–retest) of the WHIIRS was conducted; this study also provided correlational evidence of construct validity (Cronbach, 1971). In the second study, construct validity was examined through a convergent validity approach (Campbell & Fiske, 1959) that compared self-reports on the WHIIRS with objective data obtained from a wrist activity recorder (actigraphy). Common measures of sleep difficulty based on these recordings were compared with the self-reported insomnia measure. Although typically self-reports of sleep are not highly correlated with objective monitoring (e.g., Carskadon et al., 1976; Sateia et al., 2000), we expected a small but positive relationship between the two.

Study 1

Method

Sample

The sample consisted of 67,999 postmenopausal women participating in the WHI. The analyses involved the baseline data from 97.46% of the women in this sample who had complete information on the sleep items; these 66,269 women were enrolled in either the observational ($N = 40{,}984$) or clinical trial (CT; $N = 25{,}285$) arms of the WHI. The age range for these women was 50–79 years ($Mdn = 62$, $M = 62.07$, $SD = 7.41$). This was the same sample examined by Levine et al. (2003). A detailed discussion of the eligibility criteria and study design is provided in the Women's Health Initiative Study Group (WHISG; 1998).

Sleep Measure

As noted, the sleep disturbance items included in the WHI protocol were developed by sleep researchers consulting to the WHI Behavioral Advisory Committee (Matthews et al., 1997). Using these questions, Levine et al. (2003) developed the five-item WHIIRS. The five items are shown in the Appendix, and these items are intended to assess sleep initiation insomnia (or sleep latency), sleep maintenance insomnia, early morning awakening, and sleep quality. Sleep quality can be affected both by insomnia and by other sleep disturbances such as those related to breathing difficulties. The scale is scored as a simple sum of the items. The response categories were coded as 0 to 4, so the WHIIRS score could range from 0 to 20. A sixth item was included to assess sleep duration or quantity. This question asked, "About how many hours of sleep did you get on a typical night during the past 4 weeks?" The response categories were "10 or more hours," "9 hours," "8 hours," "7 hours," "6 hours," and "5 or less hours." These categories were coded from 0 to 5, with 0 corresponding to the most hours slept and 5 corresponding to the least hours slept.

Procedure

The WHI has a complex design that includes overlapping CTs designed to evaluate interventions related to reduced consumption of dietary fat, hormone replacement therapy, and calcium and vitamin D intake. In addition to the CTs, the WHI includes a large observational trial to be used, in part, to estimate risk indicators and new biomarkers. Detailed descriptions of the WHI have been presented in Rossouw et al. (1995) and the WHISG (1998). The relevance and importance of the WHI for psychologists have been discussed in Matthews et al. (1997) and in Appendix I of the WHISG (1998).

Most participants were recruited through population-based direct mailing campaigns targeted at age-eligible women, in conjunction with media awareness programs. To be eligible, women had to be 50 to 79 years old at initial screening, postmenopausal, likely to remain in the area for 3 years, and willing to provide written informed consent. Some major exclusion criteria were medical risks that made 3-year survival unlikely and participant characteristics associated with poor adherence and retention (e.g., substance abuse or dementia; see WHISG, 1998, for more detail). Between 1993 and 1998, the WHI invited 373,092 postmenopausal women 50 to 79 years of age to be screened for participation in a set of CTs and an observational study (OS). Of these women, 161,809 were eventually enrolled at 40 clinical centers in the United States.

The WHI screening procedures were complicated, in that eligibility in the three overlapping CTs as well as the OS was being determined. Briefly, participants were scheduled for three screening visits. At the first visit, consent was obtained. Women were given a physical examination and completed a personal information questionnaire (gathering data on such characteristics as age and race), a medications questionnaire, and an interviewer-administered questionnaire; depending on CT eligibility, some also completed a self-administered questionnaire containing the psychosocial instruments. The 10 sleep items were included in this latter set of items (the WHIIRS items were part of this set). Some women completed these questions at the second screening visit; for women in the CT, however, that visit was primarily focused on clinical activities (e.g., mammograms). The third screening visit involved a continued assessment for CT and OS eligibility. Several flowcharts detailing these visits were presented in the WHISG (1998).

As noted, all participants initially completed the sleep items at one of the screening visits. Test–retest reliability was assessed by examining the responses of a subset of the sample at two time points. Specifically, 2,887 women completed the sleep items at the intervals shown in Table 1. To allow same-day test–retest, 1,280 women completed the questionnaire twice. Approximately half of these women (55%) were enrolled in the OS versus the CT. Because OS participants were not required to return as frequently as CT women, the majority of the later retest data were obtained from women enrolled in the CT. The exception was that a greater percentage of OS women completed a sleep survey during the 91–365-day interval. Women in the OS returned for annual assessments, and the OS patients completed their retest survey an average of 274 days after the first administration (50% of these surveys were completed on or after the 296th day). CT women completed the survey earlier, an average of 256 days after the first administration (50% of these surveys were completed on or after the 259th day).

Psychometric Analyses

A resampling plan was used in conjunction with coefficient alpha to estimate the internal consistency of the sleep scale. Stability of the measure over time was assessed by test–retest correlations. Construct validity in Study 1 was supported by correlational evidence (Cronbach, 1971). The methodology followed for each of these procedures is described below.

Table 1
Test–Retest Correlations for the Women's Health Initiative Insomnia Rating Scale

	Test–retest correlation								
Statistic	Same day	2–7 days	8–14 days	15–21 days	22–30 days	31–45 days	46–90 days	91–365 days	366+ days
r	.956	.896	.844	.828	.730	.786	.771	.699	.663
N	1,280	464	243	163	164	142	105	163	163
% in clinical trial	44.3	70.0	88.1	88.3	85.4	88.0	79.0	23.3	81.6

Resampling procedure. One goal of this study was to assess the internal consistency of the WHIIRS. Usually, researchers present a single reliability coefficient, and there is no indication of the variability associated with this coefficient. Because of the large number of women involved in this study, we were able to use a resampling procedure to obtain an estimate of the reliability as well as the variance associated with the reliability estimate. Using this approach, we were able to compute a confidence interval on alpha without resorting to traditional methods that yield poor estimates when their stringent assumptions are violated (cf. Barchard & Hakstian, 1997). To investigate internal consistency, we adopted computer-intensive methods (Diaconis & Efron, 1983) to sample and resample the observed data. The use of resampling techniques has become increasingly widespread as computational power has grown over the past 20 years or so (Efron, 1982; Efron & Tibshirani, 1993; Lunneborg, 2000; Manly, 1997).

In this study, women were randomly sampled from our 66,269 participants in a way that permitted a woman to appear only once in a given sample, although each could appear in multiple samples. This particular sampling approach is known as random subsampling (Chernick, 1999). Twenty thousand random samples (resamples) were drawn so that the empirical sampling distribution obtained was based on 20,000 random samples each 150 in size. The number of women in each sample (i.e., 150) was chosen because samples of this size or larger are common in sleep research (e.g., Buysse, Reynolds, Monk, Berman, & Kupfer, 1989), and therefore our results provide an indication of the probable magnitude of the reliability to be obtained by users of the scale.

Reliability analyses. Two forms of reliability were explored. First, we investigated the internal consistency of the WHIIRS through the use of coefficient alpha (Cronbach, 1951) and the resampling methodology just described. Second, we examined test–retest correlations based on the subset of the sample that was administered the sleep scale more than once at varying time intervals. Because sleep disturbance is a dynamic phenomenon, we expected that the test–retest correlations would diminish over time rather than being stable through time.

Construct validity. Validity of the WHIIRS was assessed by examining the relationships between measures known to be related to sleep; 11 measures were available for this purpose. A 12th nonsleep item that was not expected to be related to insomnia was included to provide some assurance that the observed relationships were not explainable by method variance alone. A sleep quantity item was used to demonstrate that the insomnia measure is not equivalent to short sleep (Carskadon et al., 1976).

The 11 measures used were as follows: the short form of the Center for Epidemiologic Studies Depression Scale (CES-D; Burnam, Wells, Leake, & Landsverk, 1988), the 8 subscales of the RAND 36-Item Health Survey (RAND-36; Hays, Sherbourne, & Mazel, 1993), and 2 items from a symptom checklist. The 12th measure was the Negative Emotional Expressiveness Scale (NEE; King & Emmons, 1990).

Hypothesized relationships. Depression was chosen as an important construct for use in establishing the validity of the WHI sleep measure, because it has long been observed in the United States and elsewhere that depression and sleep disturbance covary such that as depression increases, sleep disturbance will also increase (for reviews, see, e.g., Boland & Keller, 1996; Dealberto, 1992; Hauri, 1974). The CES-D short form depression measure contains nine items, one of which is "Your sleep was restless." This item was excluded from our computation of the CES-D score so as not to artificially inflate the relationship between depression and sleep disturbance.

The RAND-36 contains the following eight subscales: emotional well-being, energy/fatigue, bodily pain, general health, social functioning, role limitations due to physical health, role limitations due to emotional health, and physical functioning. Hays and Stewart (1992) showed the relationship between six of these measures and the sleep scale developed for the Medical Outcomes Study (MOS). The RAND items are scored so that larger scores indicate better health (e.g., less bodily pain), whereas the WHI sleep items are scored so that larger scores represent poorer sleep. On the basis of the Hays and Stewart study, we expected that the correlations between the WHIIRS and the RAND scales would be negative. That is, better sleep (smaller scores) would be associated with better health (larger scores). Hays and Stewart did not present the correlations of the MOS sleep scale with the RAND emotional well-being and general health subscales; nonetheless, given the positive correlation between the RAND subscales, it seemed reasonable to hypothesize that these two scales would also be negatively related to the WHIIRS.

WHI participants were asked about two symptoms related to menopause: night sweats and hot flushes. Women were asked to indicate, on a 4-point scale (0–3), whether in the past 4 weeks the symptom did not occur (0) or the symptom occurred and was mild, moderate, or severe (1–3, respectively). We chose these two climacteric symptoms because there is evidence that vasomotor and somatic symptoms can interfere with sleep (e.g., Baker, Simpson, & Dawson, 1997; Hunter, 1992; Polo-Kantola et al., 1999). On the basis of this work, we expected to observe a positive relationship between reported climacteric symptoms and the WHIIRS.

Kripke et al. (2001) observed a U-shaped relationship between sleep duration and sleep complaints such that those sleeping fewer than 7 hr and those sleeping more than 8 hr report more sleep complaints. In addition, they reported U-shaped relationships between sleep duration and degree of obesity as well as degree of depression. Thus, we expected the relationship between sleep duration and the WHIIRS to be nonlinear.

Finally, the NEE measures conflict or ambivalence over emotional expression. That is, how ambivalent is an individual about expressing negative emotions? This is not a measure of situational factors affecting emotional expression but rather a measure of one's ambivalence in general about expressing negative emotions. It has been hypothesized that insomnia is associated with "personality types characterized by an inability to react outwardly and thus to discharge their feelings" (Kales, Caldwell, Preston, Healey, & Kales, 1976, p. 1134). In contrast, the theory behind the ambivalence construct does not assume that persons who are ambivalent cannot express negative emotions. Indeed, King and Emmons (1990) explained that "one purpose of the ambivalence construct is to distinguish between persons whose expressive styles are similar but whose underlying ambivalence differs" (p. 864). Thus, individuals can express negative emotions and be comfortable with these expressions, or they can feel ambivalent about them. A larger score indicates less ambivalence in expressing negative emotions but does not imply that the individual does not engage in these behaviors. The NEE was not expected to be related to insomnia and was selected for this reason.

The relationships between the variables discussed above and the WHIIRS were investigated through the use of Pearson correlation and trend analyses. Because the Pearson coefficient measures only linear relationships, the correlation ratio ($\hat{\eta}^2$) was used as an additional effect size measure (Kendall, 1952). The correlation ratio indicates that $\hat{\eta}^2 \times 100\%$ of the variance in the outcome variable is explained by the differences in categories of predictors. Eta squared also translates into Cohen's **f**. Cohen (1988) characterized a large effect size as .40, a medium effect size as .25, and a small effect size as .10.

Results

Internal Consistency

The reliability (α) for the entire sample of 66,269 women on the five-item WHIIRS was .786. The average reliability based on the 20,000 samples was almost identical ($\bar{\alpha} = .784$, $SD = .031$, $Mdn = .79$); with a sample size of at least 150, the observed alpha coefficient should fall between .70 and .85 approximately 99% of the time (i.e., this is the 99th percentile confidence interval). Conducting a large number of resampling studies permitted constructing a sampling distribution that included rare (i.e., extreme) events. Only 0.77% of the obtained reliability coefficients fell below .70. Conversely, 89.3% of the samples had reliability coefficients greater than or equal to .75.

Test–Retest Reliability

As described, 2,887 women were administered the sleep scale more than once at varying time intervals. As can be seen in Table 1, the test–retest correlations followed the expected pattern, with the correlation for same-day administrations being approximately .96, whereas tests separated by more than 1 year yielded a correlation of .66.

Construct Validity

Table 2 shows the correlation matrix of Pearson coefficients[2] among the WHIIRS, the CES-D (without the sleep item), the RAND-36 subscales, the climacteric symptoms, and the NEE. The diagonal elements are the reliabilities (coefficient αs). A quick perusal of Table 2 reveals that none of the correlations is large. It is desirable that the magnitude of the correlations not be too large because that would indicate that the same construct is being assessed by different measures (Cronbach, 1971).

The correlation between the CES-D and the WHIIRS may not seem large, but in fact a correlation of .29 does represent a nontrivial difference in means on the WHIIRS. The WHIIRS means increased in an almost monotonic fashion as the depression scores increased. Not surprisingly, given the large sample size, the test of the linear trend was statistically significant, $F(1, 65480) = 5{,}998.57$, $p < .0001$. Sample size aside, taking the mean WHIIRS score for women in each of the 12 CES-D response categories revealed that the WHIIRS mean in the largest CES-D category ($M = 10.3$ for Category 12) was 1.8 times that in the smallest depression category (i.e., $M = 5.7$ for Category 0, no depression). The largest category contained, however, only 8 women. The WHIIRS mean in the next-to-largest CES-D category was more than twice as large as the mean in the smallest depression category. The effect size ($\hat{\eta}^2$) obtained corresponded to a Cohen's **f** value of .31, which Cohen characterized as a medium-to-large effect size. Thus, although the correlation may have at first appeared modest, the mean differences were substantial, the means followed the expected monotonically increasing trend, and the effect size was found to be medium–large. These results provide evidence of the construct validity of the sleep measure.

The RAND-36 subscales were also correlated with the WHIIRS in the hypothesized direction. The correlations between the WHIIRS and the RAND subscales were somewhat smaller than the reported correlations between the RAND subscales and the MOS sleep disturbance scale. For example, Table 2 shows that the correlation between the WHIIRS and the energy/fatigue subscale was −.33. Hays and Stewart (1992) reported that the correlation between this subscale and the MOS sleep disturbance scale was −.44. Thus, although the correlations obtained in this study were smaller, they were of the same order of magnitude and were in the expected direction.

The RAND subscales were also linearly related to the WHIIRS. For example, as the pain scores increased (smaller values), the insomnia means also increased in a linear manner. The linear trend was statistically significant ($p < .0001$), and Cohen's **f** value was 0.273. The scores for all of the RAND subscales increased linearly as the WHIIRS means increased; that is, the health measure decreased as sleep disturbance increased. The values of Cohen's **f** for the remaining subscales were as follows: emotional well-being, .357; energy/fatigue, .350; general health, .273; social functioning, .248; role limitations due to physical health, .226; role limitations due to emotional health, .218; and physical functioning, .214. These effect sizes can be characterized as ranging between medium and large and provide support that the WHIIRS is sensitive to differences in reported physical and emotional health.

The relationships between the WHIIRS and the reported climacteric symptoms were also in the predicted direction, although the effect sizes were smaller than with the measures discussed so far. For night sweats and hot flushes, the values of Cohen's **f** were, respectively, .205 and .157. These can be characterized as medium and small-to-medium effect sizes. Although these effects are not as large as with the previously discussed measures, sleep complaints were again monotonically related to increased symptoms.

The relationship between the WHIIRS and the sleep duration item indicated that larger insomnia scores were associated with fewer hours of sleep. The correlation ratio ($\hat{\eta}^2 = .149$, $\mathbf{f} = .418$) indicated that 14.9% of the variance in the WHIIRS was explained by the differences in sleep duration categories. There was, however, a curvilinear relationship between sleep duration and the WHIIRS, although those sleeping the least reported the greatest disturbance. Women who reported sleeping 5 hr or less had WHIIRS scores that were almost twice as large ($M = 11.3$, $SD = 5.4$) as those of either women sleeping 7 hr ($M = 5.9$, $SD = 3.8$) or women sleeping 10 hr or more ($M = 6.2$, $SD = 4.6$).

[2] Pearson correlation coefficients are reported here. Because of possible concerns regarding the distribution of the items, Spearman coefficients were also computed and found to be essentially the same. The average difference between the two coefficients was .002, and the largest difference was .02.

Table 2

Correlations Among the Women's Health Initiative Insomnia Rating Scale (WHIIRS), Sleep Duration, and Other Constructs

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. WHIIRS	(.786)													
2. Sleep duration	.344	—												
3. CES-D (without sleep item)	.292	.127	(.650)											
4. Emotional well-being	−.334	−.147	−.688	(.819)										
5. Energy/fatigue	−.329	−.099	−.440	.563	(.878)									
6. Bodily pain	−.263	−.092	−.247	.280	.476	(.814)								
7. General health	−.262	−.080	−.285	.388	.557	.494	(.784)							
8. Social functioning	−.237	−.095	−.464	.486	.475	.490	.410	(.771)						
9. Role limitations due to physical health	−.220	−.067	−.244	.269	.493	.643	.477	.524	(.843)					
10. Role limitations due to emotional health	−.213	−.095	−.521	.538	.398	.229	.266	.482	.322	(.759)				
11. Physical functioning	−.201	−.063	−.192	.218	.469	.569	.537	.389	.564	.212	(.889)			
12. Night sweats	.152	.077	.139	−.148	−.129	−.121	−.114	−.112	−.077	−.103	−.075	—		
13. Hot flushes	.201	.074	.161	−.172	−.166	−.154	−.149	−.145	−.115	−.127	−.114	.666	—	
14. NEE	.034	−.001	.067	−.118	−.044	−.028	−.044	−.046	−.018	−.054	−.012	.051	.038	(.641)

Note. Reliabilities appear in parentheses along the diagonal. CES-D = Center for Epidemiologic Studies Depression Scale; NEE = Negative Emotional Expressiveness Scale.

Those reporting 8 or 9 hr of sleep had, on average, the least sleep disturbance (M = 5.0, SD = 3.5).

Finally, the NEE was chosen a priori as an instrument that should not be related to sleep complaints. This expectation was realized.

Discussion

The WHIIRS was found to have very good short-term test–retest reliability and acceptable internal consistency. Relative to the usual practice of presenting one reliability coefficient, the resampling procedure that we used had the advantage of providing an empirical sampling distribution from which probabilities of obtaining different outcomes could be estimated. For example, we are very certain in stating that the internal consistency will fall at or between .70 and .85 approximately 99% of the time. Similarly, there is approximately a 60% chance of obtaining a reliability around the mean value, that is, between .77 and .81. Researchers can expect to obtain a value of coefficient alpha in these ranges, with the most likely value being around .79. Reliabilities below .59 or above .88 would be a cause for concern. If the reliability falls below or above these values, respectively, it is quite possible that the data were coded incorrectly. A very low reliability likely indicates that the items were not all coded in the same direction; the sleep quality item was probably not properly coded.

Validity was assessed by examining correlations of the WHIIRS with measures of related constructs. All correlations and trends (i.e., linear or quadratic) were in the expected direction, and the WHIIRS properly showed an almost zero correlation with a measure with which it was predicted to have no relationship. The almost zero correlation of the WHIIRS and the NEE allowed us to discount the possibility that the observed relationships between the WHIIRS and the other variables were due to method variance alone. We did not have data to conduct a full multitrait, multimethod matrix, and so we cannot make a more general statement about trait validity.

The correlations observed in this study are similar in magnitude to those observed in studies involving different sleep scales. For example, Mitchell and Woods (1996) reported that in the 1st year of their study, the correlation between vasomotor symptoms and insomnia was .29. This was somewhat larger than the relationships observed in the WHI data between insomnia symptoms and either night sweats (r = .20) or hot flushes (r = .15). These results indicate that, in these two samples, the relationship between sleep and vasomotor symptoms was not large. In contrast, stronger relationships were observed by Polo-Kantola et al. (1999), who reported correlations between self-reported sleep disturbance and self-reported vasomotor symptoms of .53 for hot flushes and .58 for night sweats. Interestingly, the self-reported vasomotor symptoms in their study were not related to indicators of sleep quality as measured by polysomnography (PSG) in a sleep laboratory (these measures included latency, efficiency, number of awakenings, and sleep duration). From a validity perspective, it is important to keep distinct the difference between objective and subjective sleep disturbance. Clearly, the sleep instrument used by Polo-Kantola et al. was measuring subjective sleep disturbance, and it is likely that all self-reported sleep disturbance measures mostly tap into the subjective dimension, thus accounting for the small correlations between objective and subjective sleep indices. This issue is discussed further in Study 2.

As noted above, Hays and Stewart (1992) reported correlations between the RAND-36 subscales and the MOS sleep disturbance scale that were somewhat larger than those observed between the RAND subscales and the WHIIRS. There is, of course, variability in the correlations with the RAND subscales observed across studies. For example, the Epworth Sleepiness Scale (Johns, 1991) measures daytime sleepiness and has been shown to be related to sleep disorders. Bennett, Barbour, Langford, Stradling, and Davies (1999) reported a correlation between the 36-item Short Form Health Survey of the Medical Outcomes Study (SF-36) energy/vitality subscale and the Epworth scale of −.47, whereas the correlation between the role limitations due to physical problems subscale and the Epworth scale was −.37. In contrast, Akashiba et al. (2002) reported correlations between the same variables of −.07 and −.13, respectively. The correlations between these RAND subscales and the WHIIRS were −.33 and −.22, indicating that they fall in the range of observed correlations of sleep scales with the RAND-36 and SF-36 subscales.

The Pittsburgh Sleep Quality Index (PSQI; Buysse et al., 1989) is currently the most widely cited sleep questionnaire (cf. Levine et al., 2003). The PSQI assesses sleep quality during the previous month using 18 self-rated items and 5 items rated by a bed partner or roommate. The final PSQI score is based only on the self-rated items and is composed of seven components. Buysse et al. (1989) reported an overall coefficient alpha of .83. This value is well within the range that should be obtained when using the WHIIRS. Buysse et al. also reported that test–retest reliability after 1 to 265 days ($M = 28.2$ days) was .85. This figure is somewhat higher than the correlation noted for the WHIIRS in the 22–30-day retest period. Their distribution of number of days until retest was, however, highly skewed to the right, as indicated by a mean of 28.2 days versus the 265-day range between retests. Thus, most of Buysse et al.'s patients completed the questionnaire in many fewer days than 28.2. The test–retest correlations they reported are similar to the WHIIRS correlations for retest periods shorter than 22 days.

The correlations between the WHIIRS and the CES-D short form were .29 (without the sleep item) and .33 (with the sleep item). These correlations were smaller than those reported between the PSQI and the CES-D 20-item form, which ranged from .50 to .65 (e.g., Carpenter & Andrykowski, 1998; Ionescu, Driver, Heon, Flanagan, & Shapiro, 2001; McCurry & Teri, 1995). Differences might be due to the WHI having used the short form. The CES-D long form contains several items that are similar to those on the PSQI. These items can inflate the correlation and make it more difficult to separate whether sleep disturbance or depression is being measured. The CES-D sleep item is of course one of the overlapping items, but there are others. For example, the PSQI asks, "How much of a problem has it been for you to keep up enough enthusiasm to get things done?" This item is probably highly correlated with the CES-D items "I felt that everything I did was an effort" and "I could not get 'going.' " In any event, the PSQI exhibits larger correlations with the CES-D than does the WHIIRS. Whether it is better for sleep scales to be more or less correlated with a depression measure is an open question.

In summary, the results from this study provide evidence of construct validity. The next study examined the validity of the WHIIRS using different approaches.

Study 2

This study examined the WHIIRS against objective measures of sleep. Although there is often a poor relationship between objective and subjective sleep measures (e.g., Carskadon et al., 1976), construct validity would be supported by a demonstration that the WHIIRS yielded higher scores for those with the greatest sleep difficulties as measured by actigraphy (i.e., a portable method of measuring activity that can be used to estimate sleep). Three common measures of overall sleep difficulty were used in this study: sleep latency (time between laying down to sleep and actually falling asleep), an indicator of initiation insomnia; sleep efficiency (percentage of time in bed spent sleeping); and wake after sleep onset (WASO), or wake within sleep, which is an indicator of maintenance insomnia. Sleep latency was measured by wrist activity recorder (actigraph) from lights out to falling asleep. Sleep efficiency, measured with data both from the actigraph and from sleep logs, was computed as percentage of time in bed actually spent sleeping. WASO was measured by wrist activity recorder, and we expected insomnia scores to increase as minutes of WASO increased. Insomnia is often defined as a sleep-onset latency greater than 30 min and an accompanying sleep efficiency lower than 85% (Morin et al., 1999; Spielman, Saskin, & Thorpy, 1987). Thus, we expected that those meeting this definition of insomnia would have larger WHIIRS scores than those not falling into this category.

Method

Sample

A sample of 459 women enrolled in the observational arm of the WHI were recruited to participate in a sleep study.[3] The mean age of participants was 67.71 years. In recruiting participants, emphasis was placed on attracting women who had reported abnormal sleep duration (either 6 hr or less or more than 8 hr). Average duration of sleep per night among the participants was 6 hr 47 min, as compared with approximately 6 hr 36 min for the sample in Study 1. Although sleep duration was not less for participants in this study, the distribution was different: 43.4% of this sample reported 6 hr or less of sleep, and 8% reported more than 8 hr of sleep. In Study 1, 35.4% of women reported 6 hr or less of sleep, and 4.4% reported more than 8 hr. Thus, there was some success in obtaining a sample with a larger percentage of women with sleep abnormalities.

Procedure

Each of the 459 women wore an Actillume (Ambulatory Monitoring, Ardsley, NY) wrist activity recorder to monitor her sleep and wakefulness for a week both at home and while active in her community. Women also provided daily records of sleep–wake patterns, including daily bedtime and wake time. Each morning participants completed a log indicating the following: time lights were extinguished for sleep, time of final awakening after night's sleep, how long it took to fall asleep, and duration of sleep.

[3] This is a longitudinal study of the effects of sleep disturbance on morbidity and mortality. Baseline data from this study were used here to evaluate the construct validity of the WHIIRS.

Whereas PSG remains the standard laboratory method for assessing physiological correlates of insomnia, actigraphy offers a less expensive and more flexible alternative that also permits 24-hr recording of activity (Ancoli-Israel, 2000; Richards, 2002). The actigraph monitor used in this study, the Actillume, incorporates a photometer and a linear accelerometer along with a microprocessor and memory. The photometer is useful because it detects lights out as well as first morning light.

The monitors were mounted on padded straps worn on the wrist for continuous monitoring of activity and illumination exposure. Actillumes were initialized to record activity every minute; the activity measurement was proportional to bidirectional accelerations in the ulnar axis integrated over time. Previous research demonstrated that the Actillume reliably discriminated between sleep and wakefulness (Ancoli-Israel, Clopton, Klauber, Fell, & Mason, 1997; Mason & Kripke, 1995; Matsumoto et al., 1998). In a study that analyzed home recordings of postmenopausal women (age range: 51 to 77 years), Jean-Louis, Kripke, Cole, Assmus, and Langer (2001) found that the correlation between actigraphy and PSG was .90. Minute-by-minute agreement rates were 85% and 89% when the 24-hr recordings were scored with the rules of Webster, Kripke, Messin, Mullaney, and Wyborney (1982). They concluded the Actillume could reliably monitor sleep and wakefulness in community-residing older people. Other sleep researchers have reported similar results (e.g., Ancoli-Israel, 2000; Kushida et al., 2001).

In-bed time was obtained primarily from the Actillume illumination data, which usually indicated lights-out times accurate to 1 min. Sleep logs were used to supplement these data and to help interpret ambiguous light records, such as when someone went to bed with the lights on or turned out the lights to watch television (which can be very dim) without intending to go to sleep. Sleep–wake was inferred through an algorithm validated against home electroencephalographic PSG (Jean-Louis et al., 2001). Sleep efficiency and WASO were computed with information from the Actillume; these data were also used in computing sleep latency, but in-bed start time was obtained from both the Actillume and the sleep logs (this method has also been suggested by Kushida et al., 2001). Activity data were averaged over seven in-bed and six out-of-bed intervals; "in-bed" interval usually meant at night with lights out. On the last day of the activity recording, participants completed the WHIIRS and other quality of life instruments, including the CES-D.

Analyses

One-way analysis of variance was used to test the null hypothesis that the population means in the insomnia and noninsomnia groups were equal. These two groups were formed by applying the definition of insomnia discussed above (i.e., less than 85% sleep efficiency and latency greater than 30 min). Efficiency was computed from the actigraph and the sleep log data. Pearson zero-order correlations were used to demonstrate the degree of linear relationship between objective and subjective measures. Trend analyses were conducted to investigate the form (both linear and nonlinear) of the relationship between the WHIIRS and WASO as well as the CES-D. The correlation ratio was used as an effect size measure.

Finally, although the WHIIRS was not intended as a screening instrument by the WHI, but rather as a predictor of clinical and behavioral variables as well as an outcome, we nonetheless investigated its properties as a screening tool. We used four standard measures of accuracy to assess the predictive accuracy of the WHIIRS. Because the WHI insomnia scale does not yield dichotomous scores, it is necessary to define a score on the WHIIRS that will separate those with "insomnia" and those without. This cutpoint is required before the sensitivity (the proportion of women who truly have insomnia who were correctly classified as such) and the specificity (the proportion of women who truly do not have insomnia who were also correctly classified) can be computed. We created a receiver operating characteristic (ROC) curve (Hanley & McNeil, 1982) by plotting sensitivity against 1 − specificity so that these measures of accuracy could be examined over the range of the WHIIRS, and thus a cutpoint could be chosen that maximized specificity while keeping sensitivity above .50. The predictive value of any insomnia test also depends on the prevalence of insomnia in the population tested. A lower prevalence implies that a test must be more specific to be useful. Lower prevalence also indicates that an individual is more likely to be incorrectly identified as having insomnia. For a given specificity and sensitivity, the predictive value of a positive test (i.e., the proportion of respondents classified as having insomnia who actually have insomnia) increases as the prevalence increases. Similarly, the predictive value of a negative test (i.e., the proportion of respondents classified as not having insomnia who actually do not have insomnia) decreases as prevalence increases.

Results

Average number of minutes in bed, as measured by the actigraph, was 476.23 (*SD* = 51.22); average number of minutes of in-bed sleep was 359.82 (*SD* = 54.48; mean hours = 6.00, *SD* = 0.91). Self-reported average number of hours slept was 6.72 (*SD* = 0.99). Average sleep latency was 28.91 min (*SD* = 21.18), average WASO was 87.50 min (*SD* = 35.39), and average sleep efficiency was 75.77% (*SD* = 8.42%).

The mean WHIIRS score in Study 2 was 7.26 (*SD* = 4.92), as compared with 6.61 (*SD* = 4.45) for the sample in Study 1; thus, women in Study 2 reported somewhat more difficulty sleeping than the normative sample. The internal consistency (α) of the WHIIRS in this study was .794, a value very close to the mean of the sampling distribution in Study 1.

Using the common definition of insomnia discussed above, we compared two groups to examine whether the WHIIRS could distinguish between women with and without insomnia. The first group contained women with a latency greater than 30 min and an efficiency less than 85% (the "insomnia" group), and the second group comprised everyone else. As expected, the insomnia group reported greater sleep difficulties. The mean WHIIRS score in the insomnia group was 9.08 (*SD* = 5.58, n = 100), and the mean in the other group was 6.76 (*SD* = 4.53, n = 322). The difference between these means was about 0.48 of the pooled standard deviation and was found to be statistically significant, $F(1, 141.8) = 14.36$, $p < .001$.

The insomnia criteria of sleep latency greater than 30 min and sleep efficiency less than 85% are commonly used in the literature. Some researchers may, however, be concerned that these criteria are not as applicable to an older age group because, in this group, trouble staying asleep and early awakening are more common than trouble falling asleep. Therefore, it might be reasonable to explore using the 80% criterion of low sleep efficiency along with a latency greater than 30 min. We also used these criteria, and the results were similar to the analysis using the 85% criterion; the mean WHIIRS score in the insomnia group was 9.52 (*SD* = 5.18, n = 68), and the mean in the other group was 6.88 (*SD* = 4.72, n = 354). The difference between these means was about 0.55 of the pooled standard deviation and was again found to be statistically significant, $F(1, 420) = 17.30$, $p < .001$.

Trend analysis indicated a significant linear relationship between the WHIIRS and WASO, as measured by the actigraph, $F(1, 401) = 18.03$, $p < .0001$, $\hat{\eta}^2 = .094$, $\mathbf{f} = .322$. The deviation from linearity was not statistically significant (p = .23). Thus, as hypothesized, there was a general trend for insomnia score to increase as minutes of WASO increased.

Table 3 shows the correlations between the sleep measures based on the Actillume recordings (WASO, latency, efficiency, and duration) and the WHIIRS, the five self-report items that composed the insomnia scale, the sleep duration item, and the CES-D. As expected, the correlations were not large, but they were in the predicted directions and were statistically significant. In general, the results indicate that the objective measures correlated most highly with the WHIIRS items that were intended to tap the same facet of the insomnia construct. For example, WASO was most highly correlated with waking up several times at night (Item 2) and with total WHIIRS score. Similarly, for sleep latency the largest correlation was with trouble falling asleep. Finally, the correlations of sleep efficiency with waking up several times at night and the WHIIRS had the largest absolute magnitudes. The correlation between sleep efficiency and the WHIIRS was negative because a larger efficiency value implies better sleep, whereas a lower score on the WHIIRS indicates better sleep. Lack of efficiency is the primary indicator of maintenance insomnia, which is represented by Items 2 and 3 in the Appendix. The correlation between sleep duration as measured by self-report and by actigraph was, of course, also expected. The correlation was negative because the self-reported sleep item is scored so that a higher number indicates less sleep, whereas the actigraph reported minutes of sleep. Taken together, the relationships between the activity measurements and those of the WHI insomnia scale indicated that the WHIIRS was tapping into the insomnia construct.

As in Study 1, the CES-D (without the sleep item) was positively related to the WHIIRS ($r = .23$). A test of linearity revealed a strong linear trend, $F(1, 410) = 23.87$, $p < .0001$. After removal of the linear trend, however, there remained a significant nonlinear component, $F(11, 410) = 2.13$, $p = .017$. Both the linear and nonlinear curves were monotonically increasing and indicated that as depression increased so did insomnia score. This analysis also indicated that approximately 10% ($\hat{\eta}^2 = .104$, $\mathbf{f} = .34$) of the variance in the WHIIRS was explained by the differences in depression categories. This finding replicated the results of Study 1.

To illustrate that the WHIIRS is not simply another measure of psychological distress, consider the correlations of the CES-D with the sleep variables as measured by actigraphy. Table 3 shows that with the exception of the correlation between the CES-D and sleep latency, the correlations between the CES-D and the actigraphy sleep variables were close to zero and not statistically significant. In contrast, as noted earlier, Table 3 also indicated that the correlations between the WHIIRS and the sleep measures based on the Actillume recordings (WASO, latency, efficiency, and duration) were all statistically significant and in the predicted directions.

Finally, an ROC curve was created to address the question "What score on the WHIIRS is indicative of problematic insomnia?" In other words, who belongs in the insomnia group and who does not? Before addressing this question, we examined the area under the obtained ROC curve, which is a measure of classification accuracy. The WHIIRS was found to correctly predict those with insomnia and those without at a probability level of .65. The sensitivity and specificity of the WHIIRS were examined at various values of the insomnia scale, and Table 4 shows the consequences of four different choices. For example, choosing a score of 10 on the WHIIRS as a threshold dividing the insomnia and noninsomnia groups yielded a sensitivity of .49 and a specificity of .71. The predictive value of a positive test was .24, and the predictive value of a negative test was .88. As the threshold was decreased to a WHIIRS score of 7, the sensitivity increased to .68 and the specificity decreased to .53, the predictive value of a positive test changed to .22, and the predictive value of a negative test was .90. Although insomnia is relatively prevalent in the general population, especially in comparison with chronic diseases such as diabetes mellitus, it is still infrequent enough to yield many false positives. For this reason, we wanted to maximize specificity while maintaining sensitivity above the chance level. Hence, we chose 9 as the cutpoint because this yielded the largest specificity associated with sensitivity above .50. Of course, another investigator may place more weight on avoiding a different error and so would make a different choice. As shown in Table 4, the choice of a cutpoint affects the accuracy of prediction, as it does with all screening tests.

Discussion

Most important, this study has shown that differences in sleep latency, sleep efficiency, and WASO, as measured by the Actil-

Table 3
Correlations Among the Women's Health Initiative Insomnia Rating Scale (WHIIRS), Sleep Duration, Depression, and Actigraphy Measures

	Actigraphy measure[a]				
Instrument or item	CES-D	WASO	Latency	Efficiency	Sleep duration
WHIIRS	.227	.202	.143	−.200	−.001
Trouble falling asleep? (1)	.246	.122	.222	−.092	.009
Wake up several times at night? (2)	.123	.267	.089	−.227	.048
Wake up earlier than planned? (3)	.175	.093	.043	−.142	−.082
Trouble getting back to sleep? (4)	.093	.127	.040	−.155	−.007
Typical night's sleep (5)	.230	.140	.149	−.119	.021
Sleep duration	.157	−.098	.037	−.057	−.455
CES-D	—	−.014	.102	.003	−.064

Note. Numbers in parentheses refer to item designations in the Appendix. CES-D = Center for Epidemiologic Studies Depression Scale; WASO = wake after sleep onset.
[a] Significant at $p < .05$ if $|r| \geq .098$.

Table 4
Evaluation of the Women's Health Initiative Insomnia Rating Scale (WHIIRS) as a Diagnostic Test

WHIIRS cutpoint	Sensitivity	Specificity	PPV	NPV
10	.49	.71	.24	.88
9	.53	.67	.23	.88
8	.60	.60	.22	.89
7	.68	.53	.22	.90

Note. PPV = positive predictive value; NPV = negative predictive value.

lume, are reflected by corresponding differences in WHIIRS scores. These results add evidence in support of the construct validity of the WHIIRS. The results also suggest that a mean difference between groups of approximately one half of a standard deviation might be a guide to a clinically meaningful difference on the scale. Of course, this suggestion is tentative and needs replication.

Earlier we noted that self-reports are typically not highly correlated with "objective" monitoring (e.g., Carskadon et al., 1976; Polo-Kantola et al., 1999; Sateia et al., 2000). Buysse et al. (1989) reported that, except for sleep latency, there were no statistically significant correlations between the PSQI estimates of sleep disturbance and those obtained from PSG. The WHIIRS, in contrast, did show small but statistically significant correlations with sleep latency, sleep efficiency, and WASO.

If the WHIIRS is to be used as a screening test, the results of the sensitivity–specificity analysis should also be replicated to ensure that the values reported here do not vary drastically in another sample. It appears from our study that the accuracy of the WHIIRS as a diagnostic instrument, although adequate at certain cutpoints, is not as good as one would like. Perhaps accuracy would have improved if we had been able to use a clinical diagnosis of insomnia as the gold standard, as Buysse et al. did. In comparing healthy controls and patients with "definite or probable" major depression, they reported sensitivity and specificity as .896 and .865, respectively.[4] Whether using a clinical interview, rather than actigraphy, to define "poor sleepers" will improve the prediction accuracy of the WHIIRS is a question for future research. Of course, the sensitivity and specificity reported by Buysse et al. may have resulted from the use of a bimodal sample selected for having normal sleep or insomnia, whereas participants in our study were not chosen to be normal or poor sleepers. In any event, the sensitivity and specificity we observed are similar to or better[5] than the accuracy reported for many commonly used risk screening tests based on predictions from statistical models. That the predictive accuracy of the WHIIRS is similar to other tests, however, is an indication that using this type of screening measure may not provide the best classification, given the currently obtained predictive accuracies.

Persons wishing to use the WHIIRS as a screening device must carefully consider the cutpoint to be used. This choice obviously depends on whether sensitivity or specificity is most important to the purpose at hand. Given the results of this study, we believe that the cutpoint should not be greater than 10. Although specificity is probably most important for many of the instrument's possible applications, the steady decline in sensitivity above 10 is problematic for investigators trying to maintain somewhat of a balance between the two measures of accuracy. This recommendation is based on the assumption that the distribution of WHIIRS scores is fairly similar to the one in the normative sample. In that sample, approximately 20% of the respondents had scores of 10 or above (Levine et al., 2003). In Study 2, a similar percentage of the women, approximately 25%, had scores of 10 or above. As indicated, we chose a cutpoint of 9 because specificity is of greater concern, and as shown in Table 4, this cutpoint maximizes specificity while keeping sensitivity above .50. An investigator with a very different distribution of scores may need to choose another threshold to maintain acceptable levels of accuracy.

Women in this study completed sleep logs daily for a week. It is possible that completing the sleep logs each day heightened women's attention to the qualities of their nightly sleep and that, without this sensitization, the WHIIRS would not have been correlated with the objective measures. Future research is needed to address this possible threat to the external validity of the present study.

In summary, Study 2 provides evidence of the construct validity of the WHIIRS. It also indicates that the WHIIRS can possibly be used as a screening measure.

Conclusion

Again, most important, this study provides compelling evidence that differences in sleep latency, sleep efficiency, and WASO, as measured by the Actillume, were reflected by corresponding differences in WHIIRS scores. As in other studies, the correlations of the self-report and "objective" measures were not large. Nonetheless, we found that the WHIIRS was sensitive to group differences (e.g., those with and without insomnia, as defined by objective measures). The correlations between the objective measures and the WHIIRS and the components of the WHIIRS were in the predicted directions and made intuitive sense. For example, WASO was more highly correlated with waking up several times at night than, say, with trouble falling asleep. This finding was supportive of construct validity in that both of the former measures are indicators of maintenance insomnia, whereas the latter item is a measure of latency, with which it was most highly correlated. This again supported construct validity, because both are indicators of initiation insomnia. Validity was also supported by the correlations with other measures (e.g., the CES-D and the RAND-36) that were in the predicted directions. Reliability was found to

[4] Fichtenberg and colleagues (Fichtenberg, Putnam, Mann, Zafonte, & Miller, 2001; Fichtenberg, Zafonte, Putnam, Mann, & Millard, 2002) reported that sensitivity and specificity to insomnia of a PSQI global score above 8 were 93% and 100%, respectively. These results are flawed, however, because the insomnia group was defined through sleep logs for which questions were very similar to those of the PSQI.

[5] For example, the Gail et al. (1989) model of breast cancer risk prediction has been used for setting Food and Drug Administration (FDA) guidelines for tamoxifen use. This model was studied in a large U.S. cohort of nurses (the Harvard Nurse's Health Study); when evaluated at the cutpoint corresponding to the FDA guidelines for tamoxifen chemoprevention, it was found to have a sensitivity of .44 and a specificity of .66 (Rockhill, Spiegelman, Byrne, Hunter, & Colditz, 2001).

be acceptable to very good; short-term test–retest reliability was very good, and internal consistency was acceptable. As a whole, these findings provide support for the construct validity and reliability of the WHIIRS. As noted by Cronbach and Meehl (1955), construct validation is an ongoing process. As such, the results presented here can be regarded as the first steps in that process.

The primary limitation of this study is that only older women were included. Also, these women self-selected to participate in the WHI. Future research needs to explore the reliability and validity of the WHIIRS among younger women as well as among men. In addition, further research is needed to replicate the sensitivity and specificity of the instrument at different cutpoints.

It is also important to note that the 5-item WHIIRS was embedded in 10 sleep items and was not administered independently of the other 5 items. The WHIIRS items and the other items were not highly correlated, and we have no reason to expect the performance of the instrument to change when administered without the other items; however, this remains an empirical question.

Despite these limitations, taken together, the results of the present studies provide evidence in support of the reliability and construct validity of the WHIIRS. The results also suggest how large a difference between groups is needed to be clinically meaningful (approximately 0.5 *SD*) and which score could be used as a cutpoint between those with and without insomnia. Thus, we conclude that the WHIIRS is now ready for testing outside of the WHI.

References

Akashiba, T., Kawahara, S., Akahoshi, T., Omori, C., Saito, O., Majima, T., & Horie, T. (2002). Relationship between quality of life and mood or depression in patients with severe obstructive sleep apnea syndrome. *Chest, 122,* 861–865.

American Academy of Sleep Medicine. (1997). *International classification of sleep disorders: Diagnostic and coding manual, revised.* Rochester, MN: Author.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Ancoli-Israel, S. (2000). Actigraphy. In M. H. Kryger, T. Roth, & W. C. Dement (Eds.), *Principles and practice of sleep medicine* (3rd ed., pp. 1295–1301). New York: Saunders.

Ancoli-Israel, S., Clopton, P., Klauber, M. R., Fell, R., & Mason, W. (1997). Use of wrist activity for monitoring sleep/wake in demented nursing-home patients. *Sleep, 20,* 24–27.

Baker, A., Simpson, S., & Dawson, D. (1997). Sleep disruption and mood changes associated with menopause. *Journal of Psychosomatic Research, 43,* 359–369.

Barchard, K. A., & Hakstian, A. R. (1997). The robustness of confidence intervals for coefficient alpha under violation of the assumption of essential parallelism. *Multivariate Behavioural Research, 32,* 169–191.

Bennett, L. S., Barbour, C., Langford, B., Stradling, J. R., & Davies, R. J. O. (1999). Health status in obstructive sleep apnea, relationship with sleep fragmentation and daytime sleepiness, and effects of continuous positive airway pressure treatment. *American Journal of Respiratory and Critical Care Medicine, 159,* 1884–1890.

Boland, R. J., & Keller, M. B. (1996). Outcome studies of depression in adulthood. In K. I. Shulman & T. Mauricio (Eds.), *Mood disorders across the life span* (pp. 217–250). New York: Wiley.

Burnam, M. A., Wells, K. B., Leake, B., & Landsverk, J. (1988). Development of a brief screening instrument for detecting depressive disorders. *Medical Care, 26,* 775–789.

Buysse, D. J., Reynolds, C. F., Monk, T. H., Berman, S. R., & Kupfer, D. J. (1989). The Pittsburgh Sleep Quality Index—A new instrument for psychiatric practice and research. *Psychiatry Research, 28,* 193–213.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81–105.

Carpenter, J. S., & Andrykowski, M. A. (1998). Psychometric evaluation of the Pittsburgh Sleep Quality Index. *Journal of Psychosomatic Research, 45,* 5–13.

Carskadon, M. A., Dement, W. C., Mitler, M. M., Guilleminault, C., Zarcone, V. P., & Spiegel, R. (1976). Self-reports versus sleep: Laboratory findings in 122 drug-free subjects with complaints of chronic insomnia. *American Journal of Psychiatry, 133,* 1382–1388.

Chernick, M. R. (1999). *Bootstrap methods: A practitioner's guide.* New York: Wiley.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). Washington, DC: American Council on Education.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 53,* 281–302.

Dealberto, M. J. (1992). Les troubles du sommeil en psychiatrie: Aspects epidemiologiques [Epidemiology of sleep disorders in psychiatric conditions]. *Encephale, 18,* 331–340.

Diaconis, P., & Efron, B. (1983). Computer-intensive methods in statistics. *Scientific American, 248,* 116–130.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans.* Philadelphia: Society for Industrial and Applied Mathematics.

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap.* New York: Chapman & Hall.

Fichtenberg, N. L., Putnam, S. H., Mann, N. R., Zafonte, R. D., & Miller, A. E. (2001). Insomnia screening in postacute traumatic brain injury: Utility and validity of the Pittsburgh Sleep Quality Index. *American Journal of Physical Medicine & Rehabilitation, 80,* 339–345.

Fichtenberg, N. L., Zafonte, R. D., Putnam, S., Mann, N. R., & Millard, A. E. (2002). Insomnia in a postacute brain injury sample. *Brain Injury, 16,* 197–206.

Floyd, J. A., Medler, S. M., Ager, J. W., & Janisse, J. J. (2000). Age-related changes in initiation and maintenance of sleep: A meta-analysis. *Research in Nursing and Health, 23,* 106–117.

Foley, D. J., Monjan, A. A., Izmirlian, G., Hays, J. C., & Blazer, D. G. (1999). Incidence and remission of insomnia among elderly adults in a biracial cohort. *Sleep, 22*(Suppl. 2), S373–S378.

Ford, D. E., & Cooper-Patrick, L. (2001). Sleep disturbances and mood disorders: An epidemiologic perspective. *Depression and Anxiety, 14,* 3–6.

Ford, D. E., & Kamerow, D. B. (1989). Epidemiologic study of sleep disturbances and psychiatric disorders: An opportunity for prevention? *Journal of the American Medical Association, 262,* 1479–1484.

Gail, M. H., Brinton, L. A., Byar, D. P., Corle, D. K., Green, S. B., Schairer, C., & Mulvihill, J. J. (1989). Projecting individualized probabilities of developing breast cancer for White females who are being examined annually. *Journal of the National Cancer Institute, 81,* 1879–1886.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143,* 29–36.

Hauri, P. (1974). Sleep in depression. *Psychiatric Annals, 4,* 45–62.

Hays, R. D., Sherbourne, C. D., & Mazel, R. M. (1993). The RAND 36-Item Health Survey 1.0. *Health Economics, 2,* 217–227.

Hays, R. D., & Stewart, A. L. (1992). Sleep measures. In A. L. Stewart & J. E. Ware Jr. (Eds.), *Measuring functioning and well-being: The Med-*

ical Outcomes Study approach (pp. 235–259). Durham, NC: Duke University Press.

Hunter, M. (1992). The south-east England longitudinal study of the climacteric and postmenopause. *Maturitas, 14,* 117–126.

Ionescu, D., Driver, H. S., Heon, E., Flanagan, J., & Shapiro, C. M. (2001). Sleep and daytime sleepiness in retinitis pigmentosa patients. *Journal of Sleep Research, 10,* 329–335.

Jean-Louis, G., Kripke, D. F., Cole, R. J., Assmus, J. D., & Langer, R. D. (2001). Sleep detection with an accelerometer actigraph: Comparisons with polysomnography. *Physiological Behavior, 72,* 21–28.

Johns, M. W. (1991). A new method for measuring daytime sleepiness: The Epworth Sleepiness Scale. *Sleep, 14,* 540–545.

Kales, A., Caldwell, A. B., Preston, T. A., Healey, S., & Kales, J. D. (1976). Personality patterns in insomnia: Theoretical implications. *Archives of General Psychiatry, 33,* 1128–1134.

Kendall, M. G. (1952). *The advanced theory of statistics* (Vol. 1, 5th ed.). London: Charles Griffin.

King, L. A., & Emmons, R. A. (1990). Conflict over emotional expression: Psychological and physical correlates. *Journal of Personality and Social Psychology, 58,* 864–877.

Kripke, D. F., Brunner, R., Freeman, R., Hendrix, S., Jackson, R. D., Masaki, K., & Carter, R. A. (2001). Sleep complaints of postmenopausal women. *Clinical Journal of Women's Health, 1,* 244–252.

Kushida, C. A., Chang, A., Gadkary, C., Guilleminault, C., Carrillo, O., & Dement, W. C. (2001). Comparison of actigraphic, polysomnographic, and subjective assessment of sleep parameters in sleep-disordered patients. *Sleep Medicine, 2,* 389–396.

Levine, D. W., Kaplan, R. M., Kripke, D. F., Bowen, D. J., Naughton, M. J., & Shumaker, S. A. (2003). Factor structure and measurement invariance of the Women's Health Initiative Insomnia Rating Scale. *Psychological Assessment, 15,* 123–136.

Lunneborg, C. E. (2000). *Data analysis by resampling: Concepts and applications.* Pacific Grove, CA: Duxbury.

Maggi, S., Langlois, J. A., Minicuci, N., Grigoletto, F., Pavan, M., Foley, D. J., & Enzi, G. (1998). Sleep complaints in community-dwelling older persons: Prevalence, associated factors, and reported causes. *Journal of the American Geriatric Society, 46,* 161–168.

Manly, B. F. J. (1997). *Randomization, bootstrap and Monte Carlo methods in biology.* London: Chapman & Hall.

Mason, W., & Kripke, D. F. (1995). Comparison of the Actillume and EEG for identifying total sleep time and wake after sleep onset. *Sleep Research, 24,* 482.

Matsumoto, M., Miyagishi, T., Sack, R. L., Hughes, R. J., Blood, M. L., & Lewy, A. J. (1998). Evaluation of the Actillume wrist actigraphy monitor in the detection of sleeping and waking. *Psychiatry and Clinical Neuroscience, 52,* 160–161.

Matthews, K. A., Shumaker, S. A., Bowen, D. J., Langer, R. D., Hunt, J. R., Kaplan, R. M., et al. (1997). Women's Health Initiative—Why now? What is it? What's new? *American Psychologist, 52,* 101–116.

McCurry, S. M., & Teri, L. (1995). Sleep disturbance in elderly caregivers of dementia patients. *Clinical Gerontologist, 16,* 51–65.

Mellinger, G. D., Balter, M. B., & Uhlenhuth, E. H. (1985). Insomnia and its treatment: Prevalence and correlates. *Archives of General Psychiatry, 42,* 225–232.

Mitchell, E. S., & Woods, N. F. (1996). Symptom experiences of midlife women: Observations from the Seattle Midlife Women's Health Study. *Maturitas, 25,* 1–10.

Morin, C. M., Hauri, P. J., Espie, C. A., Spielman, A. J., Buysse, D. J., & Bootzin, R. R. (1999). Nonpharmacologic treatment of chronic insomnia. *Sleep, 22,* 1134–1156.

Ohayon, M. M. (2002). Epidemiology of insomnia: What we know and what we still need to learn. *Sleep Medicine Reviews, 6,* 97–111.

Ohayon, M. M., Caulet, M., & Lemoine, P. (1998). Comorbidity of mental and insomnia disorders in the general population. *Comprehensive Psychiatry, 39,* 185–197.

Polo-Kantola, P., Erkkola, R., Irjala, K., Helenius, H., Pullinen, S., & Polo, O. (1999). Climacteric symptoms and sleep quality. *Obstetrics and Gynecology, 94,* 219–224.

Richards, K. C. (2002). Actigraphy. In T. L. Lee-Chiong Jr., M. J. Sateia, & M. A. Carskadon (Eds.), *Sleep medicine* (pp. 689–696). Philadelphia: Hanley & Belfus.

Rockhill, B., Spiegelman, D., Byrne, C., Hunter, D. J., & Colditz, G. A. (2001). Validation of the Gail et al. model of breast cancer risk: Implications for chemoprevention. *Journal of the National Cancer Institute, 93,* 358–366.

Rossouw, J. E., Finnegan, C. P., Harlan, W. R., Pinn, V. W., Clifford, C., & McGowan, J. A. (1995). The evolution of the Women's Health Initiative: Perspectives from the NIH. *Journal of the American Medical Women's Association, 50,* 50–55.

Sateia, M. J. (2002). Epidemiology, consequences, and evaluation of insomnia. In T. L. Lee-Chiong Jr., M. J. Sateia, & M. A. Carskadon (Eds.), *Sleep medicine* (pp. 151–160). Philadelphia: Hanley & Belfus.

Sateia, M. J., Doghramjii, K., Hauri, P. J., & Morin, C. M. (2000). Evaluation of chronic insomnia: An American Academy of Sleep Medicine review. *Sleep, 23,* 243–308.

Schwartz, S., Anderson, W., McDowell, C., Stephen, R., Cornoni-Huntley, J., Hays, J. C., & Blazer, D. (1999). Insomnia and heart disease: A review of epidemiologic studies. *Journal of Psychosomatic Research, 47,* 313–333.

Spielman, A. J., Saskin, P., & Thorpy, M. J. (1987). Treatment of chronic insomnia by restriction of time in bed. *Sleep, 10,* 45–56.

Webster, J. B., Kripke, D. F., Messin, S., Mullaney, D. J., & Wyborney, G. (1982). An activity-based sleep monitor system for ambulatory use. *Medical and Biological Engineering and Computing, 20,* 741–744.

Women's Health Initiative Study Group. (1998). Design of the Women's Health Initiative clinical trial and observational study. *Controlled Clinical Trials, 19,* 61–109.

World Health Organization. (1992). *ICD-10: International statistical classification of diseases and related health problems, 10th revision* (Vol. 1). Geneva, Switzerland: Author.

(*Appendix follows*)

Appendix

Women's Health Initiative Insomnia Rating Scale

These questions ask about your sleep habits. Please mark *one* of the answers for each of the following questions. Pick the answer that best describes how often you experienced the situation in the *past 4 weeks.*

	No, not in past 4 weeks	Yes, less than once a week	Yes, 1 or 2 times a week	Yes, 3 or 4 times a week	Yes, 5 or more times a week
1. Did you have trouble falling asleep?	□$_0$	□$_1$	□$_2$	□$_3$	□$_4$
2. Did you wake up several times at night?	□$_0$	□$_1$	□$_2$	□$_3$	□$_4$
3. Did you wake up earlier than you planned to?	□$_0$	□$_1$	□$_2$	□$_3$	□$_4$
4. Did you have trouble getting back to sleep after you woke up too early?	□$_0$	□$_1$	□$_2$	□$_3$	□$_4$

5. Overall, was your typical night's sleep during the past 4 weeks:

Very sound or restful	Sound or restful	Average quality	Restless	Very restless
□$_0$	□$_1$	□$_2$	□$_3$	□$_4$

Received July 10, 2001
Revision received February 4, 2003
Accepted February 12, 2003 ■