Factor Structure and Measurement Invariance of the Women’s Health Initiative Insomnia Rating Scale

Douglas W. Levine
Wake Forest University School of Medicine

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

Deborah J. Bowen
Fred Hutchinson Cancer Research Center

Michelle J. Naughton and Sally A. Shumaker
Wake Forest University School of Medicine

As part of the Women’s Health Initiative Study, the 5-item Women’s Health Initiative Insomnia Rating Scale (WHIIRS) was developed. This article summarizes the development of the scale through the use of responses from 66,269 postmenopausal women (mean age = 62.07 years, SD = 7.41 years). All women completed a 10-item questionnaire concerning sleep. A novel resampling technique was introduced as part of the data analysis. Principal-axes factor analysis without iteration and rotation to a varimax solution was conducted for 120,000 random samples of 1,000 women each. Use of this strategy led to the development of a scale with a highly stable factor structure. Structural equation modeling revealed no major differences in factor structure across age and race–ethnic groups. WHIIRS norms for race–ethnicity and age subgroups are detailed.

Sleep researchers have often lamented the lack of consistency across the various definitions of insomnia (e.g., Harvey, 2001; Ohayon, 2002; Sateia, 2002). Depending on how one groups the 84 categories of sleep and waking disturbance listed in the International Classification of Sleep Disorders (ICSD: American Academy of Sleep Medicine, 1997), approximately 37 (Harvey, 2001) to 42 (Sateia, Doghramji, Hauri, & Morin, 2000) of these categories correspond to an insomnia disorder. The matter becomes more complex when creating a concordance with the other two major classification systems: namely, the Diagnostic and Statistical Manual of Mental Disorders (4th ed.; DSM–IV; American Psychiatric Association, 1994) and the International Classification of Diseases (10th ed.; ICD-10; World Health Organization, 1992). These latter two classification systems focus on symptoms, whereas the ICSD concentrates on etiology. Underlying this difference in approach is a debate regarding the status of insomnia as a diagnosis. In other words, is insomnia merely a symptom of some underlying pathology, or is it in fact a clinical diagnosis on its own (Harvey, 2001)? Given these variations in approaches and assumptions, it is perhaps not surprising that patients classified as having insomnia by one set of criteria might be classified differently by another set of criteria (Buysse et al., 1994; Ohayon, 2002). In addition to creating discrepancies in diagnoses, this definitional complexity makes developing and validating instruments to measure insomnia difficult indeed.

As described subsequently, the purpose of the current study was to develop and evaluate a sleep disturbance scale using responses to items collected from a large sample of women. The definitional issues become relevant when assessing the validity of the items relative to the definitions of insomnia. Consider the DSM–IV’s definition of primary insomnia:

> ...a complaint of difficulty initiating or maintaining sleep or of non-restorative sleep that lasts for at least 1 month (Criterion A) and causes clinically significant distress or impairment in social, occupational, or other important areas of functioning (Criterion B). The disturbance in sleep does not occur exclusively during the course of another sleep disorder (Criterion C) or mental disorder (Criterion D) and is not due to the direct physiological effects of a substance or general medical condition (Criterion E). (American Psychiatric Association, 1994, p. 553)

Using the DSM–IV (or the ICD-10) criteria requires evaluating the presence of a set of symptoms rather than focusing on etiology. A diagnosis made with the ICSD, in contrast, necessitates specifying an underlying pathology (Harvey, 2001). The nosologies also differ as to whether they specify criteria regarding the chronicity and severity of insomnia symptoms (Harvey, 2001; Ohayon, 2002). The ICD-10 requires a patient to experience sleep disturbance at least 3 nights per week before an insomnia diagnosis is considered. The DSM–IV and the ICSD do not specify how often a complaint must occur during a week. The ICD-10 is also the only system that explicitly considers symptom severity (although the DSM–IV’s Criterion B could be considered severity). It should be...
noted, however, that there is no commonly accepted severity criterion that is either accurate or validated.

Not surprisingly, the instruments developed to assess insomnia reflect the differences in definition. In a tour de force, Sateia et al. (2000) reviewed the assessment of chronic insomnia. In their Table 6, they commented on almost 20 self-report assessment measures (mainly diaries), whereas their Table 7 included more than a dozen sleep questionnaires. These instruments ranged in length from 8 items to 863 items. Clearly, the shorter instruments could not cover the etiology in any great detail and tended to concentrate on symptoms. Sateia et al. indicated that most of these measures have been used only once. Because many of these studies involved relatively small samples, it is difficult to determine the reliability and validity of the instruments across a variety of individuals and settings. In our Discussion section in this article, the more widely used sleep instruments are reviewed in comparison with the one developed here. It is worth noting that all of the scales are measures of the intensity of insomnia symptoms that do not distinguish between primary and secondary diagnoses.

It hardly needs to be emphasized that the measurement of insomnia is of great importance because it has been estimated that 60 million Americans suffer from insomnia annually, and this number is expected to grow to 100 million by the middle of the 21st century (Chilcott & Shapiro, 1996). Epidemiologic studies often show that women and older persons are more likely to have accompanying psychological distress, somatic anxiety, major depression, and multiple health problems (Ford & Cooper-Patrick, 2001; Mellinger, Balter, & Uhlenhuth, 1985; Sateia, 2002; Sateia et al., 2000). Given the prevalence and importance of sleep disorders, it is not surprising that many clinical and observational trials now assess sleep difficulties as an essential element of quality of life. The need for a brief, reliable, stable, and well-validated measure of sleep disorders prompted the Women’s Health Initiative (WHI) to develop its own set of items in the early 1990s, at a time when there was no widely used, short, reliable, and valid scale. As stated, the goal of the current study was to develop and evaluate a sleep scale using responses to items collected from a large sample of the WHI participants.

The WHI 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. This 15-year study, ending in 2007, has a complex design that includes overlapping clinical trials (CTs) designed to evaluate interventions related to reduced consumption of dietary fat, hormone replacement therapy (HRT), 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. In all, 161,809 women were enrolled in the various arms of the study. Detailed descriptions of the WHI have been presented in Rossouw et al. (1995) and the Women’s Health Initiative Study Group (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).

Because of the unique database available to us, we were able to develop a short sleep scale and also conduct an extensive cross-validation of the factor structure using a novel resampling procedure. In addition, we were able to examine measurement invariance across age and race–ethnicity groups as well as replicate this invariance across multiple samples. The final scale is presented along with norms for age and race–ethnicity groups.

Method

Sample

The sample consisted of 67,999 postmenopausal women participating in the WHI. The analyses included the baseline data from 97.46% of the women in our sample who had complete information on the 10 sleep items; these 66,269 women were enrolled in either the observational (N = 40,984) or CT (N = 25,285) arms of the WHI. The age range for these women was 50–79 years (Min = 62, M = 62.07, SD = 7.41). Other demographic information collected for this sample included education, income, and marital status. The vast majority of the women had education that extended beyond high school: 20.63% had a high school diploma or less; 36.82% had some college, vocational school, or trade school; 41.73% were 4-year college graduates or postgraduates; and 0.82% were missing data on education. Household income was distributed as follows: 37.52% of women had incomes of $34,999 or below; 37.96% had incomes in the $35,000 to $74,999 range; 18.27% had incomes of $75,000 or more; and 6.24% had missing data. In terms of marital status, 4.68% of the sample had never been married; 32.13% were widowed, divorced, or separated; 62.76% were married or living in a marriagelike arrangement; and data were missing for 0.43% of the women. A detailed discussion of the WHI sample and methodology was provided in the WHISG (1998).

Sleep Measure

The sleep disturbance items included in the WHI were developed by sleep researchers consulting to the WHI Behavioral Advisory Committee (Matthews et al., 1997). The 10 items shown in Table 1 were intended to assess (in the order shown) medication use or sleeping aids, somnolence or daytime sleepiness, napping, sleep initiation insomnia or sleep latency, sleep maintenance insomnia (Items E and F), early morning awakening, snoring (an indicator of sleep-disordered breathing), perceived adequacy of sleep or sleep quality, and sleep duration or quantity. For the sleep items shown in Table 1, participants rated the frequency of sleep-related complaints over the “past 4 weeks” on a 5-point scale (coded 0 to 4). For snoring (Item H), an additional “don’t know” category was added, and more than half of the respondents used this category (50.8%). It was decided that if a respondent did not know whether she snored, then there was no subjective sleep disturbance from snoring. For these women, the “don’t know” category was recoded as a 0. Eight of the items were coded so that a larger score indicated greater sleep disturbance. Conversely, Items I and J in Table 1 were originally coded such that higher numbers indicated more sleep quality and greater sleep duration, respectively. These items were reverse coded to be consistent with the other items.

To judge whether item content reflected sleep disturbance, consider how the items match the nosologies. Respondents answered each question by thinking about how often per week, in the past 4 weeks, they experienced the situation described. Thus, “in the past 4 weeks” corresponds to the DSM–IV criterion of symptoms lasting at least 1 month. Each item measured frequency per week consistent with ICD-10 criteria, but frequency was not specified in the DSM–IV or the ICSID. Use of medications (Item A)
is not a criterion for insomnia diagnosis in either the DSM–IV or the ICD-10. Criterion E of the DSM–IV does require that the sleep disturbance not be due to a medication, yet under “Associated Features and Disorders,” the DSM–IV states that “individuals with Primary Insomnia sometimes use medications inappropriately” (American Psychiatric Association, 1994, p. 554). The ICS states reliance on medications (to the point at which they no longer are effective) as hypnogenic dependency insomnia (ICSD code 780.52-0, ICD-10 code F13.2, DSM–IV code 307.41-1). Thus, the nosologies do not specify how often a drug must be used as an aid to be considered problematic.

Item B, daytime fatigue, is an indication of the consequences of insomnia referred to in DSM–IV Criterion B and in the ICD-10. The DSM–IV also mentions that there could be impairments in the social and occupational realms but does not offer a definition of impairment or distress in social, occupational, or other areas of functioning. The WHI included only this general impairment item. Excessive daytime sleepiness is also a symptom of narcolepsy (ICSD code 347, ICD-10 code G47.4, DSM–IV code 347).

Item C, napping, is not per se a criterion listed in the DSM–IV, although it might be viewed as a consequence of insomnia. The manual notes that primary insomnia subsumes several ICSD diagnoses, one of which is “inadequate sleep hygiene” (ICSD code 307.41-1, ICD-10 codes F51.0 and T78.8, DSM–IV codes 307.42–307.47); excessive napping is one feature of this ICSD diagnosis. There was not, however, a quantitative definition of excessive. Snoring (Item H) also, is not listed as an insomnia criterion; snoring is associated with breathing-related sleep disorder (DSM–IV code 780.59, ICD-10 codes G47.3 and R06.3, ICSD codes 780.51-0–780.51-1 and 780.53-0–780.53-1).

Sateia (2002) remarked that “the accepted clinical definition of insomnia is a complaint of difficulty initiating or maintaining sleep, early awakening, poor sleep quality, or insufficient amounts of sleep” (p. 152). The remaining items (D–G, I, and J) all fit into this definition as well as with the DSM–IV criteria.

In summary, the WHI items appear to correspond to the characteristics noted in the nosologies and the literature. In addition, these characteristics are present in other sleep scales (e.g., Buysse, Reynolds, Monk, Berman, & Kupfer, 1989; Hays & Stewart, 1992). The observed correspondence with the classification systems and other scales (which are surrogates for other sleep experts) serves as an indicator of the content validity of these items (cf. Haynes, Richard, & Kubany, 1995).

Table 1
Sleep Items Used in the Women’s Health Initiative Protocol

<table>
<thead>
<tr>
<th>Item</th>
<th>Item designation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Did you take any kind of medication or alcohol at bedtime to help you sleep?</td>
<td>A</td>
</tr>
<tr>
<td>Did you fall asleep during quiet activities like reading, watching TV, or riding in a car?</td>
<td>B</td>
</tr>
<tr>
<td>Did you nap during the day?</td>
<td>C</td>
</tr>
<tr>
<td>Did you have trouble falling asleep?</td>
<td>D</td>
</tr>
<tr>
<td>Did you wake up several times at night?</td>
<td>E</td>
</tr>
<tr>
<td>Did you wake up earlier than you planned to?</td>
<td>F</td>
</tr>
<tr>
<td>Did you have trouble getting back to sleep after you woke up too early?</td>
<td>G</td>
</tr>
<tr>
<td>Did you snore?</td>
<td>H</td>
</tr>
<tr>
<td>Overall, was your typical night’s sleep during the past 4 weeks:</td>
<td>I</td>
</tr>
<tr>
<td>(0) very sound or restful, (1) sound or restful, (2) average quality, (3) restless, or (4) very restless?</td>
<td></td>
</tr>
<tr>
<td>About how many hours of sleep did you get on a typical night during the past 4 weeks?</td>
<td>J</td>
</tr>
</tbody>
</table>

Note. Response categories for Items A–H were as follows: (0) no, not in past 4 weeks; (1) yes, less than once a week; (2) yes, 1 or 2 times a week; (3) yes, 3 or 4 times a week; and (4) yes, 5 or more times a week. For Item H, an additional “don’t know” category was added. Items I and J were reverse coded so that a higher number indicates greater insomnia and fewer hours of sleep. This ordering corresponds with the other items in which higher scores indicate greater insomnia. The reverse-coded scale is presented here.

Procedure

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. Major exclusion criteria included 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, because 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 information 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 sleep items were included in this latter set of items. Some women completed these questions at the second screening visit; for women in a CT arm, 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. A set of flowcharts detailing these visits was presented in the WHISG (1998).

Psychometric Analyses

A resampling plan was used in conjunction with exploratory factor analysis (EFA) to develop and cross-validate the sleep scale. Multiple-group structural equation modeling (SEM) was used to assess measurement invariance, that is, whether the factor structure remained the same across age and race–ethnic groups. The methodology followed for each of these procedures is described below.

Resampling procedure. The goal of this study was to develop a scale with a stable factor structure that holds across different sites and study
populations. Usually, researchers report results from one EFA and sometimes also conduct a cross-validation on a subset of the original sample or on another sample. However, cross-validation is left for future studies. Because of the large number of women involved in this study, we were able to provide a detailed investigation of the stability of the scale's factor structure.

To investigate the stability of the factor structure, 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 (e.g., Efron, 1982; Efron & Tibshirani, 1993; Good, 2001; Lunneborg, 2000; Pesarin, 2001; Politis, Romano, & Wolf, 1999). In this study, 20,000 random samples (resamples) were drawn by randomly sampling 1,000 women 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).

**EFA.** As we discuss explicitly in the Results section, six different factor structures were investigated. The first set of factor analyses was conducted on all 10 sleep items. The remaining factor analyses were conducted with subsets of these items as suggested by the initial analyses. For each factor analysis, the general approach was to obtain a random sample of 1,000 different women drawn from the original sample of 66,269 women. For each random sample, we retained a summary of a measure of sampling adequacy (MSA) developed by Kaiser, Meyer, and Olkin (see Kaiser, 1970b; Kaiser & Rice, 1974). The MSA is one indicator of the psychometric adequacy of the sample correlation matrix. The value of MSA lies between 0 and 1, with a higher value indicating greater sampling adequacy. Kaiser and Rice (1974) characterized values of the MSA as follows: 9 = marvelous, 8 = meritorious, .7 = middling, .6 = mediocre, .5 = miserable, and less than .5 = unacceptable.

For each random sample, we also retained a summary of the factor structure yielded by a principal-axes factor analysis without iteration3 using a varimax rotation on the resulting factors. The number of factors retained was determined with Kaiser's rule (i.e., retaining factors with associated eigenvalues > 1). For a single-factor analysis, items were designated as belonging to the factor on which the item loaded most highly. This procedure was repeated 20,000 times, each time sampling 1,000 distinct women from the original sample. The results of the 20,000 different factor analyses were used to investigate the stability of the solutions. If the factor structure were stable, only a few patterns should appear frequently out of the 20,000 analyses. If the scale were poorly defined, the results would have a multitude of different patterns each occurring relatively infrequently.

The sample size of 1,000 for each factor analysis was chosen as the number that most researchers would agree should yield a stable factor solution with 10 items. Many rules of thumb (e.g., 10 cases per variable) would suggest that much smaller sample sizes are needed, but we chose the upper limit (suggested by Connery & Lee, 1992, p. 217) to allay concerns that the different factor structures observed from sample to sample were due to insufficient sample sizes. Coincidentally, for bootstrap resampling, Lunneborg (2000, p. 97) suggested that with a large population the sample size should ideally be "no more than 1% of the population. More realistically,. . .the large population shortcut is appropriate if N is at least 20 times the size of n" (i.e., n < 5% of the population). Because a sample of 1,000 is 1.51% of 66,269, a sample size of 1,000 seemed reasonable from the point of view of both factor analysis and random resampling.

**Structural equation models.** Multiple-group SEM was used to compare the equivalence of the factor structure across race–ethnic and age groups in 20 cross-validation studies. Assessment of equivalence, or measurement invariance, is important because if the measurement structure differs across groups, unambiguous interpretation of observed group differences is not possible owing to the confounding effects of differences in measurement. The first step in determining the comparability of the models across groups was to arrive at a baseline model that fit the data for each group. If the same model could be fit to each group, the model was said to have "form invariance" (i.e., the same paths and same fixed and free parameters).

Because measurement invariance is a matter of degree, if form invariance was observed we then examined whether the factor loadings, or slopes, were equivalent across groups (i.e., "factor invariance"). For example, if women are divided into three age groups, 50–59, 60–69, and 70–79 years, we can test the null hypothesis of equality of slopes across age groups: $H_0: \Lambda^{50–59} = \Lambda^{60–69} = \Lambda^{70–79}$, where $\Lambda^i$ is the vector of regression weights for age group $i$.

Because of the nested nature of the models (i.e., the model with constraints on the slopes is a subset of the baseline model), the difference in the chi-square values for the baseline model and the constrained model can be used to test the equality hypothesis. If the hypothesis of equal factor loadings was not rejected, we proceeded to a nested series of even more restrictive equivalence constraints by placing these constraints on the intercepts, means of the latent variable, the variance–covariance matrix of the errors, and finally the latent variable's variance (Bollen, 1989).

The substantive interpretation of these tests is provided in the presentation of the results, but one example is given here. The latent insomnia variable is presumed free of measurement error, so in the Platonic sense (Levine, 1994), each person has a "true" value of insomnia. People with the same true value of insomnia experience the same difficulties sleeping, and people with different true values have different experiences. If the slopes or the intercepts linking the latent variable to the observed variables differ across age groups, then individuals of different ages with the same true degree of insomnia will differ systematically on the observed indicators of insomnia. This scenario indicates that a score on the observed scale has different meanings for different groups; this is the essence of differential item functioning (Holland & Wainer, 1993).

---

3 In this procedure, the diagonal of the correlation matrix remains unchanged. The resulting eigenvalues associated with the principal components are interpreted as the amount of variance accounted for by each component. Using Kaiser's rule here makes intuitive sense because any eigenvalue less than 1 indicates that the original diagonal of the correlation matrix (i.e., a variance of 1) does better than the new factor resulting from transformation of the correlation matrix (this was not the rationale given for this "rule" by Kaiser, 1970; Douglas W. Levine was taught this reasoning by Ingram Olkin). Although there are concerns about using Kaiser's rule to determine the number of factors, as there are with all methods of this type, these concerns do not seem to be particularly salient in this study. Given the large number of factor analyses and the relatively small number of resulting factors, it is difficult to maintain that use of Kaiser's rule resulted in too many factors having been extracted.

The component method used here is very popular; it does differ from other factor models, however, although the models yield results whose differences are often not of practical concern (Velicer & Jackson, 1990). To allay any misgivings regarding the analyses reported, we conducted a smaller resampling study using principal-axes factoring with iteration; here the elements of the correlation matrix's main diagonal were replaced with squared multiple correlations as the initial estimates of the communalities. This smaller study resulted in all 2,000 resamplings showing one-factor solutions, the same result obtained with the component method.

In a final substudy, we examined the effect on our findings, if any, of using a nonorthogonal rotation. The 10 sleep items were factor analyzed through principal-axes factoring with iteration and a direct oblimin oblique rotation with gamma set at 0 (this yields the most oblique solution and is equivalent to quartimin; see Harman, 1967, p. 326). Two-, three-, and four-factor solutions were specified, and for each we conducted a resampling study that consisted of 2,000 resamples each 1,000 in size. The results of these 6,000 analyses supported those reported here.
Because there are at least 100 formal hypothesis tests of equality of parameters across age and race groups in the 20 studies, we also present a somewhat loose “global index” of invariance to provide a quick overview of the degree of equivalence observed across all of the studies. The baseline model consisted of five indicators of the latent insomnia variable, namely, Items D, E, F, G, and I. In addition, the covariances between some of the errors were estimated: namely, $D \leftrightarrow I$, $E \leftrightarrow F \leftrightarrow G$. 4 The notation $D \leftrightarrow I + E$, for example, is read as the covariance between the errors associated with Items D and I was estimated as was the covariance between the errors associated with Items I and E.

In the baseline model, there were potentially 14 parameters per group to estimate: 4 regression coefficients (the 5th is fixed at 1), 4 covariances between the errors and 5 variances associated with the errors, and the variance associated with the latent insomnia variable. If there were only two groups, there would be 28 different parameters to estimate. If the equality constraints all held across the groups, there would be a total of 14 parameter estimates that would apply to both groups. If one equality constraint did not hold—for example, the regression coefficient for “typical night’s sleep” was not the same across the two groups—then there would be 15 parameters to estimate: the 13 parameter estimates equal across both groups and the 2 estimates for parameters that were not equal. In this example, there is no longer perfect invariance across groups, but neither is there evidence of complete inequality. This situation is termed partial measurement invariance. 4 Really this is just another example of invariance being a matter of degree, as noted above. A simple index of the degree of invariance is just the proportion of parameters that were equivalent. Thus, in the example, of the 28 parameters, 26 were equivalent (i.e., 93%). There is no hard rule as to how much partial invariance is acceptable; thus, whether this is an acceptable degree of invariance depends on the reader.

The hypotheses underlying the tests of the hierarchy of invariance described above are very stringent, in that they specify that the population parameters are exactly the same across groups. Even if the discrepancy between the model and the data is small, a large enough sample size will result in almost any model being rejected (Bollen, 1989). Because it is well known that the chi-square test of significance is sensitive to sample size, we chose a sample size for these analyses based on several considerations. Most important, because there were only 292 Native Americans in the data set, we were constrained to limit the size of each of the groups to no more than this number if the group sizes were to be kept equal. Statistical considerations also indicated that 200 cases per group is a reasonable sample size for computing multigroup models (Boomsma & Hoogland, 2001; Hoelter, 1983). Thus, in examining invariance across the groups, we decided to sample 200 women from each of the groups (1,200 women total for race and 600 total for age analyses). Reproducibility of these results was examined by cross-validating with 20 different randomly drawn samples: 10 resamples for the age analyses and another 10 for the race–ethnic analyses. Including 200 women per group, then, allowed for an adequate sample size for each analysis and also allowed for some variability in the Native American women selected in the cross-validation analyses.

We report the chi-square statistic as one measure of model fit as well as four other common fit indices: the normed chi-square ($\chi^2/df$), the comparative fit index (CFI; Bentler, 1990), the standardized root-mean-square residual (SRMR; Jöreskog & Sörbom, 1989), and the root-mean-square error of approximation (RMSEA; Browne & Cudeck, 1993; Steiger, 1999, 2000). There seems to be consensus that a normed chi-square value less than or equal to 2 represents a good fit (e.g., Bollen, 1989; Byrne, 1989; Marsh & Hocevar, 1985). For the CFI, SRMR, and RMSEA, Hu and Bentler (1998, 1999) recommended using cutoff values “close to” .95, .08, and .06, respectively.

Results

Factor Structure of the WHI Sleep Items

Six different factor structures were investigated, with the first set being conducted on all 10 sleep items. The remaining sets were conducted with subsets of these items suggested by the initial analyses. In the interest of space, not all of these analyses are reported in detail.

EFA using all 10 items. The average value of the MSA in the 20,000 studies was .77 (range: .71–.82), indicating that the correlation matrices were suitable for EFA. The 20,000 EFA studies of 1,000 women yielded two-, three-, and four-factor solutions. Three-factor solutions were by far the most common result, with 90.9% of the studies yielding a three-factor solution. In the remaining studies, 5.3% of the solutions resulted in four factors with eigenvalues greater than 1, and 3.8% of the solutions had only two factors. Because we were interested in developing a scale with a stable factor structure, it did not seem fruitful to further explore the two- and four-factor solutions.

For the samples with a three-factor solution, there were 25 different patterns of items loading on the factor associated with the largest eigenvalue (we called this “Factor 1”). Although there were 25 different patterns, more than 67% of the samples were accounted for by two patterns, namely, DEFGIJ and EFGI (letters refer to the item designation given in Table 1). These two patterns differed by only one item, namely, Item D (“Did you have trouble falling asleep?”). Among the 25 patterns, 83.34% of the samples involved some combination of only the six items DEFGIJ. From a face–content validity viewpoint, we observed that four of these items were representative of complaints associated with initiation and maintenance insomnia (i.e., chronic inability to fall asleep or remain asleep for an adequate length of time). Thus, for several reasons it made sense to further explore a scale involving these six items. 6

Analyses using Items DEFGIJ. Four scales using these items were evaluated: a six-item insomnia rating scale labeled “IRS6” (Items DEFGIJ); a five-item scale, “IRS5” (Items DEFGI); 4 As is well known, extraneous factors such as method variance, or method effect, can create a correlation between the errors (cf. Bollen, 1989, p. 232; Byrne, 1998, p. 147). Other factors such as time-specific experiences (e.g., local history effects) can also cause errors to be correlated. In fact, any variance shared across items that remains unaccounted for by their linear (in the parameters) relationships to the latent factor will result in errors being correlated. Given that it is fairly rare for a model to account for all of the variance and given that the sleep items are correlated, it would be desirable to specify covariances between all of the error terms. Because there were insufficient degrees of freedom to permit this, it was necessary, a priori, to arbitrarily choose the covariances just described.

5 Partial measurement invariance simply means that not all parameters are tested for their invariance across groups or that not all parameters are found to be equivalent across groups (Byrne, Shavelson, & Muthén, 1989). Thus, most parameters are constrained to be equal across groups, whereas some are estimated freely for each group. Models that differ across groups because, for example, additional paths or covariances are included in one group but not another can nonetheless be tested for equivalence in the parameters that are hypothesized to be equal across the groups (e.g., Byrne, 1998, pp. 266–281). 6 Items A, B, C, and H were analyzed separately because the initial analyses indicated that they did not cluster with the other items. These analyses clearly indicated that Item A (medication use) was not measuring the same construct as the other items. Nonetheless, the results did not provide strong support for a scale composed of the three items B, C, and H. Because these items did not appear to form a coherent scale, we omit analyses related to developing a scale using Items ABC.
Table 2

*Tests of Factor Invariance for Age Models Using the Women’s Health Initiative Insomnia Rating Scale*

<table>
<thead>
<tr>
<th>Study</th>
<th>Unconstrained model $H_0$ Form$^{(a)}$ equal</th>
<th>Constrained model</th>
<th>$H_0$: $\Lambda^{(b)}$ equal</th>
<th>$H_0$: $\Phi^{(b)}$ equal</th>
<th>$H_0$: $\Theta^{(b)}$ equal</th>
<th>$H_0$: $\Psi^{(b)}$ equal</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\chi^2$</td>
<td>$p$</td>
<td>$\chi^2/df$</td>
<td>CFI</td>
<td>SRMR</td>
<td>RMSEA</td>
</tr>
<tr>
<td>1</td>
<td>4.39</td>
<td>.22</td>
<td>1.46</td>
<td>.994</td>
<td>.004</td>
<td>.048</td>
</tr>
<tr>
<td>2</td>
<td>1.16</td>
<td>.76</td>
<td>0.39</td>
<td>1.000</td>
<td>.007</td>
<td>.000</td>
</tr>
<tr>
<td>3</td>
<td>3.44</td>
<td>.33</td>
<td>1.15</td>
<td>1.000</td>
<td>.006</td>
<td>.027</td>
</tr>
<tr>
<td>4</td>
<td>4.08</td>
<td>.13</td>
<td>2.04</td>
<td>.998</td>
<td>.000</td>
<td>.073</td>
</tr>
<tr>
<td>5</td>
<td>5.58</td>
<td>.13</td>
<td>1.86</td>
<td>.997</td>
<td>.008</td>
<td>.065</td>
</tr>
<tr>
<td>6</td>
<td>2.46</td>
<td>.29</td>
<td>1.23</td>
<td>1.000</td>
<td>.000</td>
<td>.035</td>
</tr>
<tr>
<td>7</td>
<td>3.97</td>
<td>.27</td>
<td>1.32</td>
<td>.999</td>
<td>.009</td>
<td>.040</td>
</tr>
<tr>
<td>8</td>
<td>4.89</td>
<td>.18</td>
<td>1.63</td>
<td>.998</td>
<td>.015</td>
<td>.056</td>
</tr>
<tr>
<td>9</td>
<td>5.76</td>
<td>.12</td>
<td>1.92</td>
<td>.997</td>
<td>.011</td>
<td>.068</td>
</tr>
<tr>
<td>10</td>
<td>3.58</td>
<td>.31</td>
<td>1.19</td>
<td>.999</td>
<td>.009</td>
<td>.031</td>
</tr>
</tbody>
</table>

Note. Boldface elements reflect partial invariance. CFI = comparative fit index; SRMR = standardized root-mean-square residual; RMSEA = root-mean-square error of approximation.

* Studies 4 and 6, $df = 2$; all others, $df = 3$.  
  ** Studies 3 and 9, $df = 6$; Study 10, $df = 7$; all others, $df = 8$.  
  * Studies 1–4, 7, and 9, $df = 8$; Studies 5, 6, and 10, $df = 7$; Study 8, $df = 6$.  
  * Studies 8 and 9, $df = 16$; Studies 1 and 10, $df = 17$; Studies 2–5 and 7, $df = 18$; Study 6, $df = 19$.  
  * Study 3, $df = 1$; all others, $df = 2$.

“IRS4,” a four-item scale (Items EFGI); and “IRS3,” a three-item scale (Items FGI). For each scale evaluated, we again conducted 20,000 factor analytic studies, and the sample size remained at 1,000 women. The results for the best of these scales, IRS5, are presented below. IRS5 was obtained by dropping Item J (number of hours of sleep) from IRS6. In IRS6, the average communality associated with Item J ($h^2 = .25$) was much smaller than the communalities associated with the other variables, the smallest of which averaged .40. The small communality for Item J was an indication that the item could be dropped from the scale.

**EFA of the IRS5 scale.** IRS5 was renamed the WHI Insomnia Rating Scale (WHIIRS) because the results indicated that it had the best combination of factor stability, average MSA value, item content, and measurement invariance (discussed below) in comparison with IRS3, IRS4, and IRS6. The WHIIRS consists of Items D, E, F, G, and I. As noted, four of these items were related to initiation insomnia, maintenance insomnia, or early morning awakening. The fifth item pertained to sleep quality, which is affected by insomnia as well as other sleep disturbances such as those related to breathing difficulties. In this set of 20,000 EFAs evaluating Items DEFGI, the average value of the MSA was .75 (range: .68–.81), 100% of the solutions had one factor, and on average 55.3% of total variance was explained by the factor. The average communalities for the variables were .407 (Item D), .483 (Item E), .601 (Item F), .660 (Item G), and .612 (Item I).

**Invariance of the Factor Structure**

Multiple-group SEM was used to compare the similarity of the factor structure across race–ethnic and age groups. The baseline model used was described above.

**Age analyses.** To evaluate the invariance hypotheses across age groups, we grouped the women into three age categories: 50–59 years, 60–69 years, and 70–79 years. The hierarchy of invariance hypotheses tested in this study was as follows: $H_{form}$, $H_\Lambda$, $H_\Theta$, $H_\Psi$, and $H_\Phi$. That is, we first examined whether the baseline models had the same form. Next, the equivalence of the slopes ($A$) relating the observed items to the insomnia latent variable was examined. The third step examined the equivalence of the intercepts ($\tau$) and the latent means ($\kappa$) across groups. The next step examined the invariance of the covariance–covariance matrix of the errors ($\Theta$). Finally, the equivalence of the variances of the latent variables ($\Phi$) was evaluated.

The results of the tests of the equality hypotheses are shown in Table 2. The italicized elements represent tests that yielded partial invariance; the others were completely invariant. Overall, the percentage of invariant elements, averaged across all 10 studies, was 96.7%. Turning to the first equality test, form invariance, Table 2 presents chi-square results and fit indices, which together show that all but two studies (Studies 4 and 6) demonstrated form invariance. Strictly speaking, in Study 6 the model also fit the data, $\chi^2(3, N = 600) = 7.76, p = .051$, but model fit was substantially improved when, for the oldest group, the covariance between the error terms associated with Item G (trouble getting back to sleep) and Item I (typical night’s sleep) was also estimated. Similarly, this same element of the covariance matrix, when estimated for the youngest group, improved the model fit for Study 4. The test statistics and fit indices for the models with partial invariance are also presented in the tables.

The chi-square difference tests between the unconstrained (baseline) model and the model constrained to have equal regression coefficients across the three age groups revealed that there was factor invariance 7 of 10 times. Thus, for these studies, the slopes linking the insomnia latent variable to the observed items were found to be equivalent across age groups. This means that, to be clear, this set of 20,000 studies was made up of new samples, different from those used to evaluate the 10-item scale. In total, 120,000 separate factor analytic studies were conducted.

IRS3 and IRS4 were also created by dropping the items with the smallest average communality.
regardless of age group, a one-unit change in insomnia led to an expected change of size $\lambda_j$ (the slope for the $j$th item) in the observed item. Perfect invariance was not observed in Studies 3, 9, and 10. In Studies 9 and 10, the 60–69 age group differed from the others in the magnitude of the slope associated with Item I; in Study 9, it was 2.4 times larger than in the other two groups, and in Study 10, it was 1.7 times larger. For Study 3, the slope estimate associated with Item I for the two youngest groups was 2.3 times that of the oldest group. Studies 3 and 9 also differed on Item E: In Study 3, the slope estimate for the two youngest groups was 1.96 times that in the oldest group; in Study 9, the slope estimate in the 60–69 age group was 2.3 times the estimate in the other groups. Although there was only partial factor invariance for these three studies, they still exhibited a substantial degree of equivalence, in that 91.6% of the slopes in the three studies exhibited age invariance. This result, considered with the complete equivalence of the factor loadings in the other seven studies, strongly suggests that the WHIIRS yielded equivalent factor loadings across age groups.

The next tests examined the question of whether the age groups responded to the sleep items in the same manner or whether some groups responded systematically higher or lower than the other groups. The tests also examined whether the mean of the latent variables differed across groups. In these analyses, the intercept terms were constrained to be equal across groups (i.e., $H_{ij}: \tau^{ij}$ are all equal, where $\tau^{ij}$ is the vector of intercepts for age group $j$). These equality constraints on the intercepts were in addition to constraining the factor loadings to be equal across groups in all studies but Studies 3, 9, and 10. In these latter 3 studies, only those slopes that were found to be equivalent across the age groups were constrained to be equivalent; the remaining few slopes were allowed to be estimated freely. The results, shown in Table 2, revealed that the null hypothesis was not rejected in 6 of the 10 studies, providing some evidence for the equality of the intercepts across age. In Studies 6, 8, and 10, nonequivalence on the intercept associated with Item I occurred, with the intercepts being larger in the youngest group than in the other two groups: 1.79, 1.72, and 1.78 versus 1.54, 1.62, and 1.50 in Studies 6, 8, and 10, respectively. In Study 5, the intercept on Item I for the two youngest groups was 1.74, and the intercept for the oldest group was 1.42.

The latent means were found to be equivalent in all studies except Studies 3 and 10. In these two studies, the mean of the oldest group was greater than the mean of the youngest group ($p < .004$), indicating greater sleep disturbance in the oldest group. Apart from these two differences, all other latent means were equivalent. In summary, the deviation from complete invariance observed among the intercepts and means does not appear so extensive as to indicate that the groups systematically differ. There is a possibility that Item I (sleep quality) is problematic, but this is discussed later.

The hypothesis that the measurement error variances and covariances were equal for all age groups was examined by placing equality constraints on the variance–covariance matrix of the errors. These constraints were in addition to those imposed in the previous tests, with the proviso that only the parameters found to be equivalent across the age groups were constrained. The chi-square difference tests shown in Table 2 revealed that the null hypothesis of equality of the variance–covariance matrix was not rejected in 6 of the 10 studies. In the 4 studies with partial invariance, there was no consistency across studies in the parameters that were not invariant. Of the six parameter estimates found to be unequal across groups, only the variance of Item F appeared in more than 1 study as nonequivalent. This occurred in Studies 9 and 10, but in the former the 60–69 age group differed from the other two, whereas in the latter the oldest group differed from the others. Again, there was no pattern in either the items involved or the groups involved. Although these 4 studies did not demonstrate 100% equivalence of the variance–covariance matrix across groups, 94.4% of the elements in the covariance matrix were found to be invariant. Thus, we believe that there is evidence for at least partial age invariance in the variance–covariance matrix of the errors.

Finally, we investigated the equality of the variance of the insomnia latent variable across age groups (i.e., $H_{ij}: \Phi^{50–59} = \Phi^{60–69} = \Phi^{70–79}$, where $\Phi^{ij}$ is the variance of the latent variable for the $j$th group). The results indicated that the null hypothesis was rejected only in Study 3. In this latter study, the variance of the insomnia latent variable was larger in the oldest group than in the others.

**Ethnic–race analyses.** The analyses presented here parallel those of the previous section. Examination of the results in Table 3 immediately reveals that there was more partial invariance than in the age analyses. The percentage of invariant elements, averaged across all 10 studies, was reduced slightly to 95.4%. This was not surprising because there were six groups instead of three, and hence many more parameters needed to be equivalent. Over the 10 studies, there were 55 inequalities out of the 1,200 parameter estimates. Despite there being relatively few inequalities, discussing each one would require too much space; thus, only those inequalities that were consistent across studies are introduced.

The chi-square statistic and all of the fit indices indicated that the 10 baseline models fit the data. This was evidence of form invariance. The chi-square difference tests between the unconstrained model and the model constrained to have equal slopes revealed that there was factor invariance 8 of 10 times. For the two studies showing partial invariance, the regression coefficient associated with the item invariance of intercepts in at least 3 of the 10 studies. The nonequivalent groups were Whites in Study 11 and Asians in Study 14.

The test of invariance of the intercepts yielded the greatest number of inequalities. All but Studies 17 and 19 showed partial invariance. There was, however, no pattern of inequalities across the studies. All race–ethnic groups, with the exception of the Native American and the “other race” groups, yielded inequalities on at least one intercept estimate in at least 2 studies. The Native American and the “other race” groups showed no inequalities of intercepts for any of the studies. Items D, E, F, and I were each associated with inequalities of intercepts in at least 3 of the 10 studies. In contrast, Item G showed no inequalities of intercepts across groups for any of the studies. As noted, there was no clear pattern of group or item inequality of intercepts across studies. There was, however, a pattern in the inequalities of the latent means across studies. Six studies had groups whose means on the insomnia latent variable differed from the White race group (the reference group). The Asian group had a lower mean (i.e., better sleep) than the White group for 5 of these studies. No other racial
or ethnic group showed any pattern, and indeed most were equivalent.

The analyses regarding the invariance of the variance–covariance matrix of errors indicated that 97.2% of the elements were equivalent. There was one clear pattern of inequalities across several studies; for Item D, Native Americans had an error variance that was about 1.6 times larger than the variance in the other groups. This pattern held across five of the studies; there were no other clear patterns.

Finally, in four studies Native Americans exhibited a somewhat larger variance in the latent variable than did the other groups (about 30% greater). In two studies, Asians had smaller variances than the other groups. There were no other patterns consistent across studies.

In summary, although presentation of these results has focused on the inequalities across age and racial groups, the vast majority of the coefficients were found to be equivalent (96.7% for age and 95.4% for race). The overall conclusion to draw from these analyses is that the scale exhibits both age and race invariance in its latent variable, showing that napping increased linearly with age. In this WHI sample, the mean score on the napping item increased in a fairly linear manner from 0.75 at 50 years of age to 1.39 at 79 years (recall that a 0 to 4 scale was used). Thus, although there was a linear increase, the mean differences were not very large, and hence the small effect size. The sleep duration item was measured on a 6-point scale, 3 indicating 7 hr of sleep and 4 indicating 6 hr of sleep (see Table 1). The effect of race–ethnicity on self-reported sleep duration indicated that Whites slept the most hours (M = 3.06, or approximately 6 hr 56 min) and African Americans slept the least (M = 3.49, or approximately 6 hr 31 min, and M = 3.51, or approximately 6 hr 29 min, respectively).

To assist in the interpretation of the norms in Table 4, we provide some additional descriptive information. The overall median was 6.0, the mode was 5.0, and the range in this sample was 0 to 20. The distribution was somewhat skewed toward the right (τ₁ = .664), indicating that more women had fewer sleep complaints. The distribution was also slightly platykurtic (τ₂ = −.069), indicating that there were fewer extreme scores than found in the tails of the normal distribution, which has a kurtosis index of 0. The cumulative distribution of scores is shown in Table 5. For example, as seen in Table 5, about 75% of the women had a WHIIRS score below 10. These norms should assist in determining where an obtained sample fits relative to the “normative population”; that is, they address the question, Is there a greater or lesser degree of insomnia in my sample relative to the WHI sample? The

---

**Table 3**

*Tests of Factor Invariance for Race–Ethnic Models for the Women’s Health Initiative Insomnia Rating Scale*

<table>
<thead>
<tr>
<th>Study</th>
<th>$\chi^2(d)$</th>
<th>$p$</th>
<th>$\chi^2(df)$</th>
<th>CFI</th>
<th>SRMR</th>
<th>RMSEA</th>
<th>$\Delta \chi^2$</th>
<th>$p$</th>
<th>$\Delta \chi^2$</th>
<th>$p$</th>
<th>$\Delta \chi^2$</th>
<th>$p$</th>
<th>$\Delta \chi^2$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>5.37</td>
<td>.50</td>
<td>0.895</td>
<td>1.00</td>
<td>.011</td>
<td>.000</td>
<td>25.67</td>
<td>.14</td>
<td>23.20</td>
<td>.18</td>
<td>53.00</td>
<td>.12</td>
<td>3.94</td>
<td>.41</td>
</tr>
<tr>
<td>12</td>
<td>8.82</td>
<td>.18</td>
<td>1.471</td>
<td>.999</td>
<td>.005</td>
<td>.048</td>
<td>21.84</td>
<td>.35</td>
<td>19.27</td>
<td>.25</td>
<td>64.07</td>
<td>.06</td>
<td>7.14</td>
<td>.13</td>
</tr>
<tr>
<td>13</td>
<td>6.73</td>
<td>.35</td>
<td>1.121</td>
<td>1.00</td>
<td>.001</td>
<td>.023</td>
<td>24.67</td>
<td>.21</td>
<td>26.92</td>
<td>.08</td>
<td>53.62</td>
<td>.15</td>
<td>4.30</td>
<td>.37</td>
</tr>
<tr>
<td>14</td>
<td>6.95</td>
<td>.33</td>
<td>1.158</td>
<td>1.00</td>
<td>.014</td>
<td>.027</td>
<td>27.27</td>
<td>.10</td>
<td>28.93</td>
<td>.07</td>
<td>53.70</td>
<td>.15</td>
<td>3.16</td>
<td>.37</td>
</tr>
<tr>
<td>15</td>
<td>10.95</td>
<td>.09</td>
<td>1.825</td>
<td>.997</td>
<td>.010</td>
<td>.064</td>
<td>26.25</td>
<td>.16</td>
<td>20.35</td>
<td>.26</td>
<td>51.17</td>
<td>.13</td>
<td>7.20</td>
<td>.07</td>
</tr>
<tr>
<td>16</td>
<td>8.65</td>
<td>.19</td>
<td>1.441</td>
<td>.999</td>
<td>.002</td>
<td>.047</td>
<td>24.07</td>
<td>.24</td>
<td>28.17</td>
<td>.06</td>
<td>58.41</td>
<td>.10</td>
<td>7.07</td>
<td>.22</td>
</tr>
<tr>
<td>18</td>
<td>6.52</td>
<td>.37</td>
<td>1.087</td>
<td>1.00</td>
<td>.003</td>
<td>.019</td>
<td>17.71</td>
<td>.61</td>
<td>29.14</td>
<td>.06</td>
<td>55.02</td>
<td>.15</td>
<td>8.41</td>
<td>.08</td>
</tr>
<tr>
<td>19</td>
<td>8.57</td>
<td>.20</td>
<td>1.429</td>
<td>.999</td>
<td>.003</td>
<td>.046</td>
<td>17.80</td>
<td>.60</td>
<td>28.80</td>
<td>.09</td>
<td>56.01</td>
<td>.09</td>
<td>9.07</td>
<td>.11</td>
</tr>
<tr>
<td>20</td>
<td>8.72</td>
<td>.19</td>
<td>1.453</td>
<td>.999</td>
<td>.001</td>
<td>.047</td>
<td>23.57</td>
<td>.26</td>
<td>28.25</td>
<td>.06</td>
<td>54.41</td>
<td>.09</td>
<td>7.68</td>
<td>.18</td>
</tr>
</tbody>
</table>

Note. Boldface elements reflect partial invariance. CFI = comparative fit index; SRMR = standardized root-mean-square residual; RMSEA = root-mean-square error of approximation.

* Studies 11 and 14, df = 19; all others, df = 20.  
* Studies 13, 16, and 20, df = 18; Studies 11, 14, and 18, df = 19; Studies 17 and 19, df = 20; Study 12, df = 16; Study 15, df = 17.  
* Studies 11 and 20, df = 42; Studies 17 and 19, df = 43; Studies 13 and 14, df = 44; Study 15, df = 41; Study 18, df = 45; Study 16, df = 46; Study 12, df = 48.  
* Studies 14 and 15, df = 3; Studies 11–13 and 18, df = 4; Studies 16, 19, and 20, df = 5; Study 17, df = 2.

---

9 The statistic $\hat{\eta}^2$ is the correlation ratio. The value $\hat{\eta}^2 = .0027$ indicated that 0.27% of the variance in the WHIIRS was explained by the differences in age groups. The statistic $\hat{s}$ is Cohen’s $f$ (Cohen, 1988), an indicator of effect size. The value $\hat{s}^2 = .0027$ translates into Cohen’s $f = .052$. Cohen defined a large effect size as .40, a medium effect size as .25, and a small effect size as .10.
The resampling approach used in this study resulted in an insomnia scale that was found to have a highly stable factor structure. SEM indicated substantial equivalence across age and race–ethnic groups. The results showed a high degree of consistency across the 10 age studies and suggest that it is possible for a researcher to find measurement invariance on form, slopes, intercepts, latent means, variance–covariance matrix of the errors, and variance of the latent variable across age groups. In contrast, it is unlikely that complete race invariance will also be found by an investigator. There should, however, be no systematic differences between groups. If there is partial invariance, the degree of deviation from complete invariance should be fairly minor, with only a few coefficients being unequal across groups.

Although there were no clear patterns of lack of race invariance across the various tests of hypotheses, two groups had differences worth noting. First, in five studies the Asian group had a lower latent insomnia mean than the White group. This finding indicates that those women who reported their race as Asian did not experience as much insomnia; the observed means in Table 4 also reflect this difference. Lack of invariance in latent means is not a problem because the scale should be sensitive to mean differences between groups. The latent mean difference does not indicate differential item functioning (DIF) because it does not change the fundamental relationship between the latent score and the observed score. That is, if there is invariance in the intercepts and slopes, then those sharing a given latent mean will also share the same expected sample score. In contrast, if the latent mean were the same between groups but the observed population means differed, then there is evidence of DIF as group membership affects the observed mean. This can occur when either the intercepts or the slopes differ across groups. In the case of the Asian group, there was no evidence of DIF; rather, there was evidence only of fewer self-reported difficulties sleeping. As noted, however, even though there was no pattern of inequality of intercepts across items or race–ethnic groups, it is unlikely that a researcher will observe complete invariance of intercepts across racial groups. Because there do not appear to be any systematic differences, it is impossible to predict where the inequalities will appear.

The second group difference involved Native Americans, who had an inequality on the error variance associated with Item D (i.e., sleep latency) in half of the studies. Similarly, this group exhibited a larger variance on the latent variable in 4 of the 10 studies. Recall that there were only 292 Native Americans in the sample. The cross-validation samples were each 200 in size; this sample size was approximately 70% of the total number. This indicates that there was considerable overlap in the Native American samples across cross-validation studies. For the other groups, overlap was not a concern because the next smallest groups contained 627 women, followed by 1,659 women. It may be that the appearance of a consistently larger variance was simply a case of nearly the same sample appearing in the cross-validation studies; such consistent lack of equality did not, however, arise in this group for the other parameters. These differences warrant further study because it is difficult to know whether these results indicate some lack of invariance or whether they are merely a consequence of overlap in the cross-validation samples for Native Americans.

Although there were no substantial race–ethnicity differences on the WHIIRS, sleep duration did differ across these groups. In the literature, the finding of racial differences in sleep duration is

Table 4
Norms for the Women’s Health Initiative Insomnia Rating Scale by Race–Ethnic and Age Groups

<table>
<thead>
<tr>
<th>Group</th>
<th>M</th>
<th>SD</th>
<th>No. of cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall sample</td>
<td>6.61</td>
<td>4.45</td>
<td>66,269</td>
</tr>
<tr>
<td>Native American</td>
<td>7.39</td>
<td>5.19</td>
<td>289</td>
</tr>
<tr>
<td>50–59 years</td>
<td>7.21</td>
<td>5.34</td>
<td>142</td>
</tr>
<tr>
<td>60–69 years</td>
<td>8.08</td>
<td>5.13</td>
<td>111</td>
</tr>
<tr>
<td>70–79 years</td>
<td>6.00</td>
<td>4.50</td>
<td>36</td>
</tr>
<tr>
<td>Asian or Pacific Islander</td>
<td>5.83</td>
<td>4.17</td>
<td>1,659</td>
</tr>
<tr>
<td>50–59 years</td>
<td>5.77</td>
<td>4.28</td>
<td>640</td>
</tr>
<tr>
<td>60–69 years</td>
<td>5.63</td>
<td>4.09</td>
<td>654</td>
</tr>
<tr>
<td>70–79 years</td>
<td>6.28</td>
<td>4.09</td>
<td>365</td>
</tr>
<tr>
<td>African American/Black</td>
<td>6.21</td>
<td>4.65</td>
<td>5,722</td>
</tr>
<tr>
<td>50–59 years</td>
<td>6.30</td>
<td>4.74</td>
<td>2,759</td>
</tr>
<tr>
<td>60–69 years</td>
<td>6.17</td>
<td>4.59</td>
<td>2,149</td>
</tr>
<tr>
<td>70–79 years</td>
<td>5.98</td>
<td>4.52</td>
<td>814</td>
</tr>
<tr>
<td>Hispanic/Latino</td>
<td>6.74</td>
<td>4.90</td>
<td>2,043</td>
</tr>
<tr>
<td>50–59 years</td>
<td>6.89</td>
<td>5.09</td>
<td>1,181</td>
</tr>
<tr>
<td>60–69 years</td>
<td>6.53</td>
<td>4.66</td>
<td>682</td>
</tr>
<tr>
<td>70–79 years</td>
<td>6.56</td>
<td>4.40</td>
<td>180</td>
</tr>
<tr>
<td>White</td>
<td>6.66</td>
<td>4.41</td>
<td>55,731</td>
</tr>
<tr>
<td>50–59 years</td>
<td>6.45</td>
<td>4.43</td>
<td>22,393</td>
</tr>
<tr>
<td>60–69 years</td>
<td>6.65</td>
<td>4.37</td>
<td>22,337</td>
</tr>
<tr>
<td>70–79 years</td>
<td>7.09</td>
<td>4.42</td>
<td>11,001</td>
</tr>
<tr>
<td>Other</td>
<td>6.68</td>
<td>4.60</td>
<td>627</td>
</tr>
<tr>
<td>50–59 years</td>
<td>6.52</td>
<td>4.64</td>
<td>261</td>
</tr>
<tr>
<td>60–69 years</td>
<td>6.75</td>
<td>4.60</td>
<td>255</td>
</tr>
<tr>
<td>70–79 years</td>
<td>6.87</td>
<td>4.55</td>
<td>111</td>
</tr>
</tbody>
</table>

Table 5
Cumulative Distribution of Women’s Health Initiative Insomnia Rating Scale Scores

<table>
<thead>
<tr>
<th>Score</th>
<th>Cumulative percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>5.00</td>
</tr>
<tr>
<td>1</td>
<td>12.00</td>
</tr>
<tr>
<td>2</td>
<td>19.50</td>
</tr>
<tr>
<td>3</td>
<td>27.60</td>
</tr>
<tr>
<td>4</td>
<td>36.90</td>
</tr>
<tr>
<td>5</td>
<td>46.20</td>
</tr>
<tr>
<td>6</td>
<td>55.20</td>
</tr>
<tr>
<td>7</td>
<td>62.60</td>
</tr>
<tr>
<td>8</td>
<td>69.60</td>
</tr>
<tr>
<td>9</td>
<td>75.40</td>
</tr>
<tr>
<td>10</td>
<td>80.80</td>
</tr>
<tr>
<td>11</td>
<td>85.20</td>
</tr>
<tr>
<td>12</td>
<td>88.70</td>
</tr>
<tr>
<td>13</td>
<td>91.50</td>
</tr>
<tr>
<td>14</td>
<td>93.80</td>
</tr>
<tr>
<td>15</td>
<td>95.80</td>
</tr>
<tr>
<td>16</td>
<td>97.20</td>
</tr>
<tr>
<td>17</td>
<td>98.00</td>
</tr>
<tr>
<td>18</td>
<td>98.80</td>
</tr>
<tr>
<td>19</td>
<td>99.50</td>
</tr>
</tbody>
</table>
inconsistent, with some studies suggesting that African Americans have greater sleep problems than Whites (e.g., Foley, Monjan, Izmirlian, Hays, & Blazer, 1999; Kripke et al., 2001; Whitney et al., 1998) and other studies reporting either no racial differences or differences in the opposite direction (e.g., Blazer, Hays, & Foley, 1995; Ford & Cooper-Patrick, 2001). The differences observed in this study represent a small effect size (explaining 1.9% of the variance) that may correspond to approximately a 0.5-hr difference in time asleep. Perhaps after controlling for other factors (e.g., socioeconomic status, body mass index, and household size), these differences would disappear. It is beyond the scope of this article, however, to explore racial differences other than those related to the psychometric properties of the measure, and in that regard the sleep instrument showed no important differences. For interested readers, Kripke et al. (2001) provided further results on racial differences and sleep in the WHI.

As discussed, we observed no systematic association between age and self-reported insomnia symptoms. This finding has been observed by others as well (e.g., Fichtenberg, Zafonte, Putnam, Mann, & Millard, 2002; Hajak, 2001; Katz & McHorney, 1998; Polo-Kantola et al., 1999). It may be that this lack of association was a result of all women being more than 50 years old, and thus a “restricted age range” may have attenuated a relationship between age and insomnia. Alternatively, Kripke et al. (2001) commented that national and international surveys have shown that self-reported insomnia is especially prevalent among women after menopause. In their larger WHI sample (N = 98,705), Kripke et al. found, as we did, no relationship between age and self-reported insomnia in samples of postmenopausal women. They suggested that their results were “consistent with the interpretation that insomnia is increased less by progressive aging than by menopausal status” (Kripke et al., 2001, p. 249). This suggestion is supported by studies such as that conducted by Owens and Matthews (1998). They reported that in the 3rd year of their longitudinal study, the change from premenopausal to postmenopausal status was associated with a significant increase in the number of women reporting trouble sleeping (for those not on HRT).

The WHI included a clinical trial investigating the effect of HRT on heart disease, strokes, blood clots, osteoporosis-related bone fractures, and breast and endometrial cancer. It was also anticipated that the HRT component of the WHI could provide data on the effects of menopausal symptoms and HRT on sleep. More than 27,000 women 50–79 years of age have been participating in the HRT study. At this time, however, it is unclear as to the status of these data. On May 31, 2002, the WHI Data and Safety Monitoring Board (DSMB) halted the estrogen-plus-progestin study arm because of safety concerns (Writing Group for the Women’s Health Initiative Investigators, 2002). Only women with intact uteri were randomized to this arm. The estrogen-alone arm (for women without uteri) continues to operate. Assuming that the DSMB does not detect excessive health risks in the unopposed estrogen arm, there may be future data to investigate the interrelationship among insomnia, HRT usage, and menopausal status.

**Comparison With Other Sleep Measures**

Given the prevalence and importance of sleep disorders, there has been a need for a brief sleep disorders measure that can be used in evaluating the outcomes of interventions designed to ameliorate sleep disorders (e.g., Wilcox et al., 2000) or can be used as a covariate in studies examining the many health conditions associated with sleep difficulties (e.g., Bromberger et al., 2001). Although the use of sleep questionnaires in research is common (cf. Weaver, 2001), their use as tools to assist clinicians in assessing the severity of insomnia symptoms is less frequent. Sateia (2002) observed that although questionnaires provide an excellent means of data collection in research studies, their utility in the routine clinical setting has not been well explored, and it remains unclear how much they add to diagnostic accuracy of treatment outcome in routine clinical usage. (p. 157)

This sentiment is shared by Spielman, Yang, and Glovinsky (2000), according to whom “one of the best methods for obtaining a more balanced, comprehensive overview of a complaint of persistent insomnia is to have the patient fill out retrospective questionnaires” (p. 1241). But although “questionnaires and prospective logs certainly have their role in the assessment of insomnia...it is in the face-to-face setting of the consultation that the clinician’s skills and knowledge will find full expression” (p. 1246).

Some believe that questionnaires as screening instruments would be valuable in clinical care (e.g., Fichtenberg, Putnam, Mann, Zafonte, & Millard, 2001); however, there seems to be concurrence that although questionnaires are extremely useful in research, their use is more limited in clinical settings. The WHI originally developed the sleep items to be used in its research study. We expect that others will also use the instrument primarily in research. Although the instrument might become useful as a screening measure, its value for this use requires further evaluation (see Levine et al., 2003).

Of the extant sleep instruments that have been most favored (as measured by citations in the Institute for Scientific Information’s Web of Science), the Pittsburgh Sleep Quality Index (PSQI; Buysse et al., 1989) is currently by far the most widely cited sleep questionnaire (272 citations as of this time). The next most cited instruments, the Leeds Sleep Evaluation Questionnaire and the St. Mary’s Hospital Sleep Questionnaire, have been cited almost an equal number of times (slightly less than 70), and the Sleep Questionnaire (Johns, Gay, Goodyear, & Masterton, 1971) has received 45 citations at this time.

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: subjective sleep quality (1 item), sleep latency (2 items), sleep duration (1 item), habitual sleep efficiency (3 items), sleep disturbances (9 items), use of sleeping medications (1 item), and daytime dysfunction (2 items). Seven of these 18 items correspond to 1 of the 10 WHI sleep items, and 3 of the items correspond to 1 of the 5 WHIIRS items.

The PSQI was originally tested on 148 individuals. Buysse et al. (1989) reported an overall coefficient alpha of .83; test–retest reliability after 1 to 265 days (M = 28.2 days) was .85. They further reported that the PSQI could distinguish the group of

---

10 These items sum to 19 because one item is used in two components.
“good” sleepers and the group of “poor” sleepers. Except for sleep latency, they found no statistically significant correlations between the PSQI estimates of sleep disturbance and those obtained from polysomnography.

The St. Mary’s Hospital Sleep Questionnaire (Ellis et al., 1981) was developed to assess an individual’s previous night’s sleep and was “framed with the needs of the hospital patient in mind” (p. 93). This instrument contains 14 items related to sleep quantity, sleep quality, sleep latency, and early morning awakening. No overall scoring was suggested, and each item is analyzed separately. The questionnaire was originally tested on 93 individuals, and 4-hr test–retest correlations ranged from .70 to .96. Leigh, Bird, Hindmarch, Constable, and Wright (1988) factor analyzed the items and reported four factors. Unfortunately, this analysis revealed serious problems with the factor structure of the instrument. Three of the items had communalities less than .35, and 1 of these items had no loadings greater than .12 on any factor. Leigh et al. also reported that 4 items loaded on more than one factor, making interpretation difficult. Finally, one factor had only 1 item with a loading greater than .50, indicating that this was probably not a common factor. Should investigators desire to create composite variables using the St. Mary’s instrument, these results indicate that further psychometric investigation is required before doing so.

The Leeds Sleep Evaluation Questionnaire was “designed to quantify subjective assessments of sleep during investigations into effectiveness of hypnotic drugs” (Parrott & Hindmarch, 1978, p. 325). That is, the introduction asks respondents “How would you compare getting to sleep using the medication with getting to sleep normally, i.e., without medication?” (p. 329). As with the St. Mary’s sleep questionnaire, the Leeds instrument was developed to assess the previous night’s sleep rather than sleep quality over a period of time, as is required by the diagnostic criteria. The Leeds questionnaire contains 10 items related to ease of getting to sleep, perceived quality of sleep, ease of awakening, and how the individual felt on awakening. Factor analyses yielded four factors. One factor had only 2 items loading on it; two other factors had 3 items loading on them, but 1 item loaded across the factors equally. The factor analysis was conducted on 501 completed questionnaires, but these questionnaires came from only 133 different individuals. It is not clear how the nonindependence of observations was handled.

The Sleep Questionnaire (Johns et al., 1971) was intended to assess the quality and quantity of an individual’s sleep. The results for two versions of the instrument were reported by Johns et al. The first contained 31 items, and the second contained 27 items. The instruments measured times of falling asleep and waking up, number of night awakenings, sleep duration, and sleep quality. No specific time frame is used (e.g., “How would you describe your usual sleep?”), and no method for producing an overall score was offered. The original article provided a correlation matrix of 11 of the items.

Of course, there are many other instruments, though they have not been frequently used or cited. In terms of the instruments discussed above, the WHI items are most similar to those of the PSQI. The WHIIRS and the PSQI use the same time frame (4 weeks or 1 month), and each yields a total score. Both sets of items were developed around the same time period, and their content overlaps to a large degree, although the WHIIRS is much shorter (the PSQI includes elements that were excluded from the

WHIIRS). The WHIIRS contains the subset of the WHI items related to insomnia symptoms. Other items were excluded from the WHIIRS because of psychometric considerations (e.g., medication use, snoring, and daytime fatigue), because they were ambiguous in their relationship to insomnia, or both. For instance, sleep duration was not included because of its low commonality; also, however, sleep duration is ambiguous in that there are other reasons why people have short sleep that are unrelated to insomnia. The WHIIRS was developed with a much larger sample than the PSQI and so has available norms for older women. It also has stable psychometric properties. The factor structure is highly stable, and internal consistency and test–retest reliability (see Levine et al., 2003) are comparable to the PSQI. Nonetheless, because we do not have data on both instruments, we cannot evaluate their relative performance in assessing insomnia.

Both instruments contain most of the insomnia characteristics noted in the nosologies and the literature. For the WHIIRS, we made the decision not to include daytime fatigue, an indication of the consequences of insomnia, in the final scale. There are two observations to make regarding this decision. First, it appeared from the factor analyses that the potential consequences of insomnia (e.g., daytime fatigue and napping) did not load with the symptoms of insomnia. In other words, insomnia consequences (at least those measured here) did not correlate highly with the symptom items. Investigators desiring to add a “consequences” subscale will need to develop it carefully. The factor analyses did not indicate that daytime fatigue, napping, and medication use formed a clear factor. If such a scale is developed, it should not be combined with the WHIIRS without testing, because “consequences” and “symptoms” were perceived differently by the almost 70,000 women in this study.

A second related issue pertains to the question of the validity of including consequences in the nosologies. Although it makes intuitive sense that insomnia should have evident negative effects, a clear definition of these consequences is missing. Sateia (2002, p. 155) observed that self-reports of impairments are abundant and readily obtained but that confirmatory objective evidence is less easily acquired. Gathering data in support of construct validity is more difficult if the theory does not include clear definitions and hypotheses regarding issues such as daytime consequences. Currently, impairments perceived by the patient may be regarded differently by different clinicians. This theoretical ambiguity can yield discrepancies in the evaluation of insomnia. Ohayon (2002) commented that “the diagnoses of insomnia proposed by the DSM–IV and the ICSD classifications still lack validity as demonstrated by the difficulties in achieving good inter-rater agreements” (p. 107). Thus, we are still not at a point where we can provide an unequivocal definition of insomnia and its effects. The effect of this ambiguity is that it becomes an empirical question as to which consequences, if any, should be measured by an instrument.

A last difference between the WHIIRS and other sleep measures is that it does not include a sleep duration measure, although one is available in Table 1, Item J. If a duration item is added, or if any other items, such as those related to respiratory complaints or medication use, are included in a study, they should not be combined into one scale with the WHIIRS. Recall that the factor analyses indicated that medication use, somnolence, napping, and sleep duration were not components of insomnia symptoms. It is
especially worth emphasizing that insomnia is not generally equivalent to short sleep, as evidenced in our research and that of others (e.g., Carskadon et al., 1976).

**Psychometric Issues**

Access to a substantial portion of the WHI baseline database has permitted us to use a novel resampling approach in evaluating the stability of the factor structure. The resampling strategy used here is a luxury enabled by the large number of women participating in the WHI. For researchers with access to large databases, this resampling approach can be recommended for investigating the stability of exploratory factor models. It does, however, have the disadvantage of being very computationally intensive and time consuming in that it requires reading and sorting through a large database, and this is the slowest operation for computers today. A nice advantage of the method is that it is able to detect very rare events and provides an empirical sampling distribution from which probabilities of obtaining different outcomes can be estimated.

Results of the SEM indicate that there may be a lack of age invariance on the slope and intercept estimates for Item I (typical night’s sleep). The percentages of nonequivalent elements for all items, averaged across all 10 age studies, were 4.2% of the slope estimates and 2.7% of the intercept estimates. In contrast, the percentages of nonequivalent elements for Item I, averaged across all 10 age studies, were 10% of the slopes and 13.3% of the intercepts. Thus, Item I may deserve additional psychometric scrutiny in future research regarding age invariance.

A couple of questions frequently arise regarding scales. One is whether unequal or equal weights should be applied to the items. Our position is that the items should be unweighted (i.e., equal weights). There are various reasons for this preference, including that (a) it often does not matter whether weights are applied (Wainer, 1976) and (b) weights change from sample to sample. Because only women 50–79 years of age were used in developing the WHIIRS, we do not know whether weights obtained from the WHI sample could be applied to, for example, a sample of men.

Another concern is whether an instrument results in a large amount of missing data. In general, we have found very few missing data with this scale. In the WHI sample of almost 68,000 women, only 2.5% had missing data on the 10-item sleep scale, and only 1.6% of the women were missing 1 of the 5 WHIIRS items. Thus, an investigator who finds a large amount of missing data on this scale should be concerned. Treatment of missing data is an area for further research on the WHIIRS. Missing scores on the sleep quality item may be especially unfortunate in that the other items reflect insomnia, and so the unique content represented by this item would be lost.

**Summary**

We demonstrated that the WHIIRS has a highly stable factor structure and acceptable reliability. Furthermore, the scale showed few differences across age and ethnicity groups. We have presented norms by age and ethnicity that can be used for comparative purposes or for planning new studies. The WHI items appear to correspond to the characteristics noted in the nosologies and the literature. In a separate study, Levine et al. (2003) have provided evidence of the internal consistency and longitudinal reliability and validity of the WHIIRS. In that study, validity was assessed by examining correlations of the WHIIRS with measures of related constructs. In addition, construct validity of the WHIIRS was examined by comparing responses to the WHIIRS against objective measures of sleep, and the results indicated that differences in sleep latency, sleep efficiency, and wake after sleep could be detected by the WHIIRS.

In summary, in a large sample of older women, the WHIIRS was found to be a reliable and valid scale with one stable factor. The WHIIRS is now ready for testing outside of the WHI in other populations of women and men.

**References**


WHI INSOMNIA RATING SCALE: MEASUREMENT

135


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

---

**Members of Underrepresented Groups:**
**Reviewers for Journal Manuscripts Wanted**

If you are interested in reviewing manuscripts for APA journals, the APA Publications and Communications Board would like to invite your participation. Manuscript reviewers are vital to the publications process. As a reviewer, you would gain valuable experience in publishing. The P&C Board is particularly interested in encouraging members of underrepresented groups to participate more in this process.

If you are interested in reviewing manuscripts, please write to Demarie Jackson at the address below. Please note the following important points:

- To be selected as a reviewer, you must have published articles in peer-reviewed journals. The experience of publishing provides a reviewer with the basis for preparing a thorough, objective review.
- To be selected, it is critical to be a regular reader of the five to six empirical journals that are most central to the area or journal for which you would like to review. Current knowledge of recently published research provides a reviewer with the knowledge base to evaluate a new submission within the context of existing research.
- To select the appropriate reviewers for each manuscript, the editor needs detailed information. Please include with your letter your vita. In your letter, please identify which APA journal(s) you are interested in, and describe your area of expertise. Be as specific as possible. For example, “social psychology” is not sufficient—you would need to specify “social cognition” or “attitude change” as well.
- Reviewing a manuscript takes time (1–4 hours per manuscript reviewed). If you are selected to review a manuscript, be prepared to invest the necessary time to evaluate the manuscript thoroughly.

Write to Demarie Jackson, Journals Office, American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242.