Published in final edited form as:

Psychol Assess. 2008 September ; 20(3): 206–216. doi:10.1037/a0012764.

Assessment of Generalized Anxiety Disorder Diagnostic Criteria in the National Comorbidity Survey and Virginia Adult Twin Study of Psychiatric and Substance Use Disorders

Thomas S. Kubarych, Steven H. Aggen, John M. Hettema, Kenneth S. Kendler, and Michael C. Neale

Department of Psychiatry, Virginia Commonwealth University

Abstract

The authors investigated measurement properties of the *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition*, generalized anxiety disorder (GAD) criteria in the National Comorbidity Survey and the Virginia Adult Twin Study of Psychiatric and Substance Use Disorders (VATSPSUD). The two studies used different widely used instruments. There were significant ($p < .001$) differences in measurement of GAD due to age, study, and age-study interaction on item thresholds and factor loadings of GAD, especially when different stem-probe structures of interviews were taken into account. Item thresholds were estimated to differ by as much as $-.74$ as a function of age and $.40$ as a function of study. Despite these differences, factor scores derived from symptom criteria strongly predicted categorical diagnostic outcomes based on symptom count. It is concluded that interview structure, especially the stem-probe format of structured interviews, and wording had significant effects on study findings; that future studies in psychiatric epidemiology should use common structured interviews as much as possible; and that factor scores can be used in conjunction with sum scores as cut points to retain the advantages of both dimensional and categorical classification.

Keywords

psychiatric epidemiology; generalized anxiety disorder; measurement invariance; structured interviews; validity

Psychiatric epidemiology is the study of the distribution of psychopathology in the general population, along with the risk factors that influence that distribution. In modern psychiatric research, assessments are made with specific criteria, such as those proposed in the *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, Text Revision (DSM-IV-TR*, American Psychiatric Association, 2000). A common scientific activity is to compare rates of disorder across different groups, such as men versus women, or across different populations. Although seemingly simple, such comparisons rest on a number of assumptions, and conclusions may be invalid if these assumptions are not met.

First, we must recognize that although diagnostic classification systems produce a binary affected classification versus unaffected classification, this distinction is somewhat arbitrary. Variation in the number of symptoms reported still exists, among both those who are classified

Correspondence concerning this article should be addressed to Thomas S. Kubarych, Department of Psychiatry, Virginia Commonwealth University, P.O. Box 980126, Richmond, VA 23298-1026. tkubarych@vcu.edu.

Supplemental materials: <http://dx.doi.org/10.1037/a0012764.supp>

as having the disorder and those who are not. Therefore, a latent trait perspective may prove most useful for research inquiries into the prevalence of psychiatric disorders (Krueger, Markon, Patrick, & Iacono, 2005; Pickles & Angold, 2003). Rather than counting the number of symptoms endorsed on a questionnaire or interview and then assigning a yes-or-no diagnosis based on a predetermined cutoff, users of the latent trait approach regard symptoms as imperfect measures of an unobserved dimension, called the latent trait or factor, which is hypothesized to cause the observed symptoms.

Second, to conclude that the mean of a latent trait differs between groups, we must first know that the same latent trait is measured equivalently in both groups. If a diagnosis is based on symptom count, failure of this assumption can lead to incorrect conclusions, reducing the extent to which findings can be replicated across both populations and interview instruments. This important issue has received insufficient attention in psychiatric research. For example, discrepancies in prevalence rates of common psychiatric disorders between the Epidemiological Catchment Area (ECA) study and the National Comorbidity Survey (NCS) have been reported (Regier et al., 1998; Frances, 1998). To determine whether these differences in prevalence are genuine, we must establish that the measurement instruments used assessed the same constructs in an equivalent manner. Otherwise, we may conclude that prevalence rates differ when, in fact, only the measurement of the disorder differs. Similarly, we may be at risk of concluding that rates do not differ when, in fact, they do, although measurement differences have obscured them.

Two possible sources of failure of the measurement equivalence assumption exist. One is that the measuring instruments are genuinely different in content. Even subtle alterations in the wording or the order of the questions within the two interviews may cause participants to interpret the question differently and may thus violate the assumption of measurement equivalence. Another is that differences between the samples may also invalidate the interpretation of observed differences in diagnostic rates. For example, suppose that just one symptom of a putative disorder is rarely endorsed by men. This difference at the symptom level may well lead to different rates of diagnosis in men and women, but it does not necessarily mean that the liability to the disorder as a whole actually differs between them. Should either or both of these violations of measurement equivalence be present, observed differences in prevalence rates would be biased.

Such problems are not restricted to simple comparisons of the prevalence of disorders in different samples and populations. If measurement differs between instruments or samples, we may mislead ourselves into searching for (and perhaps finding) different causes of comorbidity when, in fact, no difference in comorbidity exists. Different patterns of familial resemblance, such as a higher heritability in men than in women, may also be caused by failures of measurement invariance (Lubke, Dolan, & Neale 2004). In principle, the same mechanism may also cause different degrees of association both between disorders and their putative risk factors and between two different disorders.

Currently, although *DSM* criteria are standardized, their assessment is not. That is, there exist a variety of interview instruments that may be used to derive a *DSM* diagnosis. Therefore, both interview structure and sample differences may contribute to measurement noninvariance of *DSM* disorders. It is possible to test for instrument structure differences when we have largely overlapping instruments. Interview instruments differ somewhat in wording and structure (Spitzer & Williams, 1985; World Health Organization, 1993). Most of these interviews include a stem-probe structure, such that only participants who have responded positively to the stem item (which is normally required for most *DSM* diagnoses) are asked the remaining probe items. Some structured interviews place all such stem items at the beginning of the interview, to prevent participants from learning to answer “no” to stems and to hence avoid

multiple probes, whereas others, to improve the flow of the interview, group the stem and probe items together for each disorder. To assess the structure and prevalence of symptom criteria in the general population, it is necessary for one to take into account the stem–probe structure of these interviews (Kubarych, Aggen, Hettema, Kendler, & Neale, 2005).

Ideally, to compare the performance of two interview formats, we would assess the same participants with both interviews. There are problems with this approach, however. First, simply re-asking the same questions may change the participant's understanding of the item and may therefore yield inconsistent responses. Second, due to the stem–probe construction, large sample sizes would be needed in order to obtain a sufficient sample of participants who have responded to the probe items on both occasions.

In this article, we assess measurement equivalence between the NCS (Kessler, 2002) and the Virginia Adult Twin Study of Psychiatric and Substance Use Disorders (VATSPSUD, Kendler and Prescott, 2006). These studies used very similar *DSM* structured interviews, although the NCS interview was based on the Composite International Diagnostic Interview (World Health Organization, 1993) and the VATSPSUD was based on the Structured Clinical Interview for *Diagnostic and Statistical Manual of Mental Disorders, Third Edition, Revised* (SCID *DSM-III-R*; Spitzer and Williams, 1985). Our approach allows us to address the following questions: (a) Are the rates of generalized anxiety disorders (GAD) from these two samples directly comparable? If not, what are the main sources of difference between them? (b) Does age account for any differences in symptom reporting? (c) If there are measurement differences between the two studies, how much practical, clinical significance do these differences make?

We use item–factor modeling (described below) to assess the effects of age, study (VATSPSUD vs. NCS), and Age $\times$ Study interaction on (a) the latent trait, (b) the factor loadings, and (c) the item thresholds of the *DSM* criteria for GAD criteria. We hope that the information provided will assist in the future development of optimized and standardized instruments that would facilitate more accurate assessments of liability to GAD, which in turn would improve the quality of both research studies and clinical practice.

Method

Participants and Measures

Sample 1 consisted of 2,163 White women and girl twins from the first wave of the population-based VATSPSUD. The twins were identified through birth certificates maintained by the Virginia Department of Health Statistics. The sample was restricted to White persons because it was estimated that it would be possible to interview less than 100 twin pairs from non-White persons, too few to obtain reliable estimates of heritabilities. The response rate was 92%. The age range was 17 years to 54 years ($M = 30.1$, $SD = 7.6$). The assessment of GAD symptoms was based on the SCID interview (Spitzer & Williams, 1985). In this interview, stem items are placed adjacent to probe items for a disorder. The lifetime prevalence of *DSM-III-R* GAD (American Psychiatric Association, 1987) with a 6-month duration criteria, without hierarchy (excluding participants who met the criteria for major depression), was 6.5%.

Sample 2 consisted of a subset (women and girls) of the 8,098 participants of the NCS (Kessler, 2002). Participants were noninstitutionalized civilians in the 48 contiguous United States, aged from 15 years to 54 years (for women, $M = 33.3$, $SD = 10.6$). The response rate was 82.6%. Lifetime prevalence of *DSM-III-R* GAD without hierarchy in this sample was 6.6%. The assessment of GAD symptoms was based on the Composite International Diagnostic Interview. The version of this interview used in the NCS places all probes at the beginning of the interview. To minimize differences in participants between the two samples, only women and girl participants were used. NCS data were collected between September 1990 and February 1992.

The major demographic characteristics of the VATSPSUD and NCS are summarized in Table 1.

Procedure

Interviews for the VATSPSUD sample were conducted from January 1987 through July 1989. Approximately 10% of the interviews were conducted by telephone, primarily when participants resided outside of Virginia. Interviewers had an undergraduate degree in a behavioral science, as well as a master's degree in a clinical area or 2 years of clinical experience, and 2 weeks of training for in-person interviews. A psychiatrist reviewed each interview and made clinical diagnoses (Kendler & Prescott, 2006).

Data collected from twins are not independent. Previous studies have often dealt with this dependency by randomly selecting one twin from each twin pair. Recent studies have shown that the resulting loss of information in this approach is much more severe than any effect of even moderate twin pair resemblance (Rebollo, de Moor, Dolan, & Boomsma, 2006). We chose instead to model fully the dependency in the data (Neale, Aggen, Maes, Kubarych, & Schmitt, 2006). The twin pair correlation on the factor was .481 for monozygotic (MZ) twins and .237 for dizygotic (DZ) twins.

For the NCS, a multistage, area-probability sampling was used. NCS interviewers had an average of 5 years of interviewing experience and attended a 7-day, study-specific training program on the use of the version of the Composite International Diagnostic Interview used in the NCS (Kessler et al., 1994). Due to the closeness of the periods in which the two studies were carried out, January 1987 to July 1989 versus September 1990 to February 1992, we do not expect cohort effects to be as substantial as those of age and study.

In addition to their placement at different places in the interview discussed above, variations in wording of the stems may also have introduced differences in the samples selected to receive the probes. The VATSPSUD stem asks whether the participant has had a period of at least 1 month when he or she has felt anxious, nervous, or worried more days than not. In the NCS interview, after being asked whether they have had a period of at least 1 month of worry or anxiety in their lifetime, participants were asked whether they were worried about more than one thing at a time, about what other people might do or what might happen to others, and whether they were worried about their mental or physical health. It is thus impossible to construct an equivalent stem between the two studies.

This inconsistency leaves us with a dilemma. We cannot obtain parameter estimates for the general population without taking the skip pattern produced by the stem-probe interview format into account (Kubarych et al., 2005), yet the skip pattern differs between the two studies. Therefore, we will present the analyses for *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition*, GAD (*DSM-IV*; American Psychiatric Association, 1994) criteria both with and without taking the skip pattern into account. The *DSM-IV* criteria are a subset of the *DSM-III-R* criteria (Kubarych et al., 2005).

Statistical Model

The data for the *DSM* GAD criteria are binary items. Just as logistic and ordinal regression offer appropriate alternatives to linear regression when modeling binary or ordinal data, item factor analysis models offer appropriate alternatives to linear factor analysis when modeling binary or ordinal item responses. Structural equation modeling and item response theory (IRT) models are, in fact, variants of the more general item factor analysis framework. Both structural equation modeling and IRT approaches have been used to study measurement noninvariance. Transforming parameter estimates from structural equation modeling to IRT is straightforward;

in fact, for the binary case, factor loadings are directly interpretable as IRT discrimination parameters, and item thresholds are directly interpretable as IRT difficulty (severity) parameters (Wirth and Edwards, 2007; Takane & DeLeeuw, 1987).

In the present report, we used a single-group item factor analysis model with covariates. In this model, the covariates may affect (a) the mean of the factor (latent trait; e.g., whether the mean of the latent variable differs between the NCS and the VATSPSUD samples or whether the trait mean differs between younger and older participants), (b) the variance of the factor, (c) the factor loadings (i.e., the regressions of the items on the factor), and (d) the means of the individual items, which, for binary data, are proportional to the item thresholds. Thus, we can distinguish between changes in the factor mean and variance, which may be considered genuine effects, and changes in the factor loadings or in the item means, which are changes in the functioning of the measurement instrument. The combined data from both NCS and VATSPSUD were treated as a single sample, with study (NCS or VATSPSUD) as one of the covariates. Age and Age $\times$ Study interaction (computed by multiplying age and study) were also studied as covariates.

For binary item data, it is not possible to estimate both the factor mean and variance effects simultaneously with all the factor loading and item threshold effects, due to underidentification of the model. There are, however, two key comparisons that can be made to test for measurement noninvariance. First, we can compare the fit of a model that specifies the effects of a covariate (age or study in this case) on the factor loadings against a model that specifies the same covariate effects on the latent variance. Second, we can compare the fit of a model that specifies covariate effects on the item thresholds with one that specifies covariate effects on the latent factor mean. We also compare the combined effects of age, study, and Age $\times$ Study interaction on item thresholds with their combined effects on the latent mean to assess whether there are additional differences in thresholds due to Age $\times$ Study interaction. We can compare the combined effects of age, study, and Age $\times$ Study interaction on factor loadings with their effects on the latent variance to test for additional differences in factor loadings due to Age $\times$ Study interaction.

First, we fitted a baseline model in which none of the covariates was allowed to affect the factor means, factor variances, item thresholds, or factor loadings. Second, we fitted models in which age or study, separately; age and study, together; or age, study, and Age $\times$ Study interaction, together, were allowed to affect the mean and variance of the latent trait (liability to GAD) and compared these with the baseline model. These comparisons assess whether the latent trait mean or variance differs across age or study and do so without accounting for possible measurement noninvariance. For evidence of whether the same construct is being measured across studies or age, we fitted models in which the covariates were allowed to affect the thresholds and compared these models with their factor mean counterparts. Similarly, we fitted models in which the covariates were allowed to affect the factor loadings and compared these with covariate effects on the factor variance. Lastly, to determine which items had the largest effects on measurement differences, we compared the effect of the covariates on each item separately. A flowchart of the model fitting sequence is available as supplemental material.

Statistical modeling makes use of various criteria for choosing between different models. We present two commonly used criteria in Tables 2 through 11. The first is -2 times the logarithm of the likelihood function ($-2\ln L$). This statistic is based on the likelihood or joint probability of the data for particular parameter values; taking the logarithm and multiplying by -2 yields a statistic that is useful for model comparison. The difference between the $-2\ln L$ statistics of two nested models is, under certain regularity conditions, asymptotically distributed as chi-square, with degrees of freedom equal to the difference between the number of parameters in the two models (MacCallum, 1995). The second, Akaike information criteria (AIC), is called

an information-theoretic criterion because it emphasizes minimizing the amount of information required to express the data in the model, therefore favoring parsimonious representations of the data. Lower (more negative) values of information theoretic criteria such as AIC reflect more parsimonious models of the data (Akaike, 1987).

Results

To test for dimensionality, accounting for the systematically missing data (participants who do not endorse the stems are missing on all probes; see Kubarych et al., 2005), we conducted full information maximum likelihood (FIML) factor analysis on the combined samples for the six *DSM-IV* probes plus a stem item (seven items). The analyses were conducted in Mx (Neale, Boker, Xie, & Maes, 2002). For one- and two-factor models, $\Delta\chi^2(8, N = 6,426) = 10.47, p > .05$. We concluded that it was reasonable to treat the seven items as unidimensional, though this is not a necessary assumption.

Analysis of Probes Only

We then followed the model fitting sequence described above. A model with no effects on the factor means or variances, item thresholds, or factor loadings for any of the covariates was fit first as the baseline for comparison. We then tested for effects on the latent mean and variance of liability to GAD due to age, alone; study, alone; age and study, together; and age, study, and Age $\times$ Study interaction, together. The results for the probes only are displayed in Table 2. The effects on the latent mean (Models 2 through 5) are significant for age ($p = .001$), study, age and study, and Age $\times$ Study interaction ($p < .001$). Compared with the baseline model, which specifies no differences whatsoever between the two studies, there are significant effects on the latent variance (Models 10 through 13) for age ($p = .002$), age and study together ($p = .001$), and age, study, and Age $\times$ Study interaction ($p = .003$). These comparisons do not account for possible failures of measurement invariance.

As previously stated, the fairest comparison with which to test for invariance of item thresholds is to compare the model with the effects of the covariate(s) on the item thresholds against the model with the effects of the same covariate(s) on the latent mean. The results for the six *DSM-IV* probes, only, are given in Table 2; all comparisons for thresholds (Models 6 through 9) are highly significant ($p < .001$). These tests do indicate failures of measurement invariance. They imply that the level of the latent trait at which participants are responding positively to the items differs across the covariates. We can use information-theoretic criteria AIC, described earlier, to compare these models for parsimony. The same model with age and study, but not interaction, is most parsimonious for the *DSM-IV* GAD thresholds (AIC = -8027.26).

To test for failures of measurement invariance with respect to factor loadings, we compared the models including the effects of the covariate(s) on the factor loadings with the models including the effect of the same covariate(s) on the variance of the latent trait liability to GAD. These are Models 14 through 17 in Table 2. Study ($p = .021$) has a significant effect on the factor loadings.

The NCS sample was recruited from the 48 contiguous U.S. states, whereas the VATSPSUD sample was recruited from individuals born in Virginia. The possibility remains that the measurement noninvariance we detected was due to regional effects. We therefore performed the same comparisons restricting the NCS sample to the southeast region (which reduces the NCS sample size and, hence, the power of the test). There was still significant measurement noninvariance between the NCS and the VATSPSUD. Due to the significant differences obtained when restricting the NCS sample to the south region, we tested whether there was measurement noninvariance within the NCS by comparing the southeast region NCS sample with the rest of the NCS. Even within the NCS study, there were significant measurement

noninvariance effects on both thresholds and factor loadings. These results are available in the supplemental material for this article.

Analysis Accounting for Stem–Probe Structure

Table 3 displays the results with the full sample, taking into account the stem–probe structure of the interviews. The most striking difference is that the effects of all covariates on the factor loadings are now highly significant ($p < .001$). The effects on the latent mean are no longer significant except for the model with age, study, and Age $\times$ Study interaction. The effects on the thresholds, however, remain highly significant for all covariates. The effect of age on the latent variance is no longer significant, whereas the effect of study on the latent variance becomes significant. Overall, measurement noninvariance increases when the stems are included in the analysis. In the general population, both thresholds and loadings contribute to measurement differences between the NCS and the VATSPSUD.

Item-Level Effects

Having identified measurement noninvariance effects for study on factor loadings, age, study, and Age $\times$ Study on thresholds at the global level, we sought to determine which particular items were noninvariant. We ran the same series of model comparisons as we did in the global case, but testing the threshold and factor loading effects one item at a time, instead of all together. The results are displayed in Tables 4 through 11. In these models, a positive effect of study indicates a higher threshold or factor loading in NCS than in VATSPSUD. Similarly, positive effects of age indicate increasing thresholds or factor loadings with age.

Table 4 shows the effect sizes of age on each of the item thresholds. There is a large ($-.74$) effect for “Did you often tire easily?” A lower threshold corresponds to a higher endorsement frequency, and we expect older participants to tire more easily than do younger participants. The threshold for this item changes more than would be expected given how much the GAD factor changes with age. Failure to take into account the general tendency of older people to tire more easily would yield artificially increased diagnoses of GAD in this population. That is, older participants' tiring gives systematically different information about GAD. This diagnostic criterion does not provide the same information about GAD in older participants as it does in younger participants. The effect is statistically significant and of substantial size. The other significant effect is also large (.73 for “Were you often irritable or especially impatient”). Here, the higher threshold indicates the reverse: Older participants systematically report being less irritable or impatient than younger participants.

The two largest effects for study on thresholds (Table 5) also show the pattern of being in opposite directions: the threshold for irritability and impatience increases (higher in the NCS, .40), whereas the threshold for difficulty concentrating is lower in the NCS ($-.31$). All items show significant effects for study on their thresholds, some positive, some negative. Table 6 shows the effects on the items when both age and study are in the model, and Table 7 shows the results adding the interaction term.

Effect sizes for age on factor loadings are shown in Table 8. As with the thresholds, it is “Did you often tire easily?” and “Were you often irritable or especially impatient” that show significant effects of age on factor loadings. Tiring easily becomes more discriminating with age, whereas irritability and impatience become less discriminating. As can be seen in Table 9, four items show significant effects of study on factor loadings. Tables 10 (age and study) and 11 (age, study, and Age $\times$ Study) give the results for combinations of age and study.

Clinical Significance

The difficulty of determining the clinical significance of discrepancies in prevalence in epidemiological studies has been widely discussed in the literature (Regier et al., 1998; Frances, 1998; Muthén, 1996; Rodebaugh, Woods, Heimberg, Liebowitz & Schneier, 2006). Full *DSM-IV* diagnostic criteria are available as supplemental material. We do not have data in the VATSPSUD interview for *DSM-IV* Criterion B (the person finds it difficult to control the worry) and E (the symptoms cause impairment in functioning) and do not deal with hierarchy—participants were not excluded if they met the criteria for major depression. A *DSM-IV* diagnosis for GAD cannot be obtained without at least three of the six criteria in this study, which are listed under Section C in *DSM-IV*. We assessed the clinical significance of measurement noninvariance by cross-tabulating sum scores of three or more against factor scores uncorrected for measurement noninvariance and factor scores corrected for measurement noninvariance.

To best examine agreement between factor scores and sum scores, we chose the factor score threshold that yielded the same proportion of factor scores as the symptom criteria (e.g., because 18% of participants had sum scores of 3 or more, we chose the 82nd percentile of the factor scores as the cutoff). The agreement between factor scores and sum scores was examined before and after correcting for measurement noninvariance. In the full sample, without correcting for measurement noninvariance, 157 participants are potentially misclassified; only 103 participants, or 34% less, are potentially misclassified when factor scores are corrected for measurement noninvariance. We then performed the same analyses separately for the NCS and VATSPSUD. For the VATSPSUD, without correcting for measurement effects, 135 of 2,163 participants (about 6%) were potentially misclassified; when corrected factor scores were used, only 89 out of 2,163 (about 4%) were potentially misclassified. For the NCS, 14 out of 4,263 participants (less than 1%) were potentially misclassified with uncorrected factor scores, versus 14 with corrected factor scores. The greater disagreement between the factor scores and the sum scores in the VATSPSUD than in the NCS is highly significant, $\chi^2(1, N = 6,426) = 130.36$. This difference may be due to the different locations of the stems, which were all at the beginning of the interview in the NCS, but not in the VATSPSUD. The results with factor scores corrected for measurement noninvariance are shown in Table 12.

The measurement noninvariance effects of age in this study tended to be counterbalancing: the strong effect of age on the threshold “Do you tire easily?” (−.73) is counterbalanced by the strong positive effect of age on irritability (.74). It is still worth inquiring, however, whether misclassification is correlated with age. To determine whether participants in the tails of the age distribution are preferentially misclassified, we computed biserial correlations between factor scores (corrected and uncorrected) and the absolute value of age deviation from the mean. There was very little correlation between age and misclassification (biserial correlation = −.048 for corrected factor scores; biserial correlation = −.064 for uncorrected factor scores).

Even though the samples differ in their agreement between diagnoses and factor scores, in both, the agreement is very good, especially when correcting for measurement noninvariance. Therefore, use of corrected factor scores for research provides maximum clinical validity. Conversely, if a clinician were to use factor scores for diagnosis, a cutoff could be used that would provide diagnoses close to *DSM*. A major advantage of the factor score approach is that it provides quantitative information about differences among cases and noncases. This alternate metric may prove especially useful in evaluating treatment or prevention effects.

Discussion

In this study, we detected differences in the measurement of a common psychiatric disorder between two widely published population-based samples. Although we cannot establish the

clinical significance of these findings, our analyses indicated that factor scores derived from symptom criteria strongly predicted diagnostic outcome based on sum score—that is, the participants who would be considered candidates for a diagnosis based on sum scores would be almost all the same participants who would have been selected on the basis of factor scores, especially in the NCS. This is an important result from both research and clinical perspectives because factor scores are measured on a continuous scale. From a research perspective, statistical power is much greater when a continuous variable is used instead of a dichotomized case–noncase variable. From a clinical perspective, a continuous variable can provide information about the effects of intervention on a participant's percentile standing on the latent variable. Thus, one might hypothetically find that a given intervention has changed a participant's standing on the latent trait from the 99th percentile to the 95th percentile, 80th percentile, or 50th percentile.

This result, indicating that comparison of rates of *DSM-IV* GAD diagnoses across the two samples is not substantially impacted by measurement noninvariance, may or may not generalize to other disorders; to comparisons across different groups, such as men and women; or even to different definitions of GAD. Therefore, caution is warranted for any comparison across studies or groups until the effects of measurement noninvariance have been investigated. Similarly, any comparison of rates of GAD across ages should be interpreted with caution, in either longitudinal or cross-sectional studies. In practice, the clinician may consider tiredness to be a more significant symptom in a younger patient and irritability or impatience to be more significant in an older patient.

Older individuals frequently report that they tire more easily, for reasons (e.g., physical fitness) that are likely not due to changes in liability to GAD. All other things being equal, an older individual who has responded positively to tiring easily is likely to have a lower liability to GAD than a younger individual. However, this item remains salient for diagnostic purposes at the older ages; in fact, it discriminates slightly better among older than among younger participants. The factor loading increases with age by a different degree, according to which covariates are included as moderators: by .38 with age alone (Table 8), by .28 with both age and study included as covariates (Table 10), and by .57 with age, study, and Age $\times$ Study interaction all used as covariates (Table 11). We expect older patients, for example, to tire more easily. It seems less clear why older participants should be less irritable and impatient and, at the same time, more keyed up and on edge. The results with respect to study are more troubling and less easily understood. Why, for example, should there be such a big difference in thresholds for the difficulty concentrating item?

There are many possible explanations for the measurement differences in these studies. First, their samples differ demographically, as summarized in Table 1. A limitation of the VATSPSUD is its lack of ethnic minorities in the sample. As stated above, this exclusion occurred because VATSPSUD researchers estimated that the minority sample would be too small to obtain reliable estimates of genetic and environmental variance components. Second, the studies used different sampling techniques. VATSPSUD researchers interviewed as many twins as possible from a database maintained by the Virginia Department of Health Statistics. The NCS used a stratified, multistage, area probability sample.

A third possible source of measurement differences is that the placement of the stems differs between the NCS interview, in which all stems are grouped at the beginning of the interview, and the VATSPSUD, in which they are not. Possibly, some participants in the VATSPSUD have learned that by saying “no” to the stems, they can avoid having to answer a long list of additional questions. The interviews also differ slightly in their wording of the items, as documented in the supplemental material. These are, however, only hypotheses, and further research is needed to resolve between these alternatives. Because, for their cost-effectiveness

and statistical power, studies depend in part on accurate measurement (Rao & Gu, 2002), we believe that these issues deserve a high priority. To avoid the possibilities of differences in wording of items or placement of stems causing measurement noninvariance, we recommend that future studies in psychiatric epidemiology use a commonly agreed on interview as much as possible. The alternative would seem to be developing ever more complex statistical methods for adjusting for differences between studies, which may, in turn, require larger sample sizes and make comparisons between studies or samples impossible for all but the most statistically sophisticated researchers.

Moving on to the probes, the NCS interview prefacing the probes with the sentence “When you were worried or anxious, were you also ...” followed by short, bullet-type questions, whereas the VATSPSUD has no prefacing sentence and uses full sentences rather than bullets for the probes. Also, the qualifier *often* appears in five of the six VATSPSUD probes but in none of the NCS probes. It is quite possible that these minor differences in wording caused differences in the endorsement frequencies. This possibility, however, may be masked by differences in selection due to the different wording of the stem items.

We found greater measurement noninvariance when the skip pattern produced by the stem–probe format of the structured interview was taken into account than when analyses were performed only on the probes. This increase highlights the importance of taking the stem–probe structure into account when the objective of a study is to obtain parameter estimates that are valid for the general population.

Our findings are especially important with respect to current research priorities (Cuthbert, 2005). The National Institute of Mental Health (NIMH) is currently giving priority to research aimed at understanding the pathophysiology of mental disorders. Measurement issues such as those described in this article are crucial for these goals. Advances in fields such as neuroscience and genetics, as in any science, depend on accurate, valid measurement. Identifying reliable genetic associations, biomarkers, and neural circuits are facilitated by detailed and accurately measured phenotypes. Some genes involved in psychiatric disorders are likely to be relevant to a spectrum of disorders rather than a single *DSM* diagnosis. Krueger et al. (2002), for example, found that heritability of an externalizing factor was 81%. Identifying genes related to the externalizing spectrum will be relevant to all the disorders in the spectrum. Improved measurements may greatly assist our understanding of how different genes, environments, and neural circuits interact to produce psychopathology. We are currently investigating whether other disorders, such as major depression, have measurement noninvariance across studies, sex, and age. These issues are particularly salient in light of the notoriously high comorbidity among psychiatric disorders. As Cuthbert (2005) points out, the common strategy of selecting participants with one and only one disorder may be biased because such participants may be less severely affected than are those who are diagnosed with more than one disorder.

As stated above, the fact that we did not find many participants misclassified due to measurement noninvariance could be due to the lack of a *DSM-IV* diagnosis in these data and does not preclude participants being misclassified in other disorders or due to other covariates. At the same time, the close relationship between the quasi-diagnostic criteria and the estimated factor scores implies that factor scores provide valuable supplemental information. These data should, by providing greater statistical power, assist in the identification of high risk populations and of risk factors that increase liability to GAD. The clinical significance of measurement noninvariance deserves further study. It is possible, for example, that measurement noninvariance across sex may be responsible for reported sex differences in certain disorders. It is also possible that some cut points for a given disorder generate higher rates of measurement noninvariance misclassification than do others.

Limitations

Our analyses were limited to women and girls and might not generalize to men and boys. These findings also pertain to the associated symptom criteria, not the binary diagnosis of GAD. The participants in the VATSPSUD are 100% twins, whereas only 2% of the NCS participants would be expected to be twins. Twins do not differ significantly from nontwins, however. It was impossible to construct an equivalent stem between the two studies, and the results might differ if the interviews had used the same stems. The NCS used a multistage, area probability sampling method, whereas the VATSPSUD participants were identified through birth certificates maintained by the Virginia Department of Health Statistics.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by National Institutes of Health Grants MH-65322, MH-40828, and MH/AA/DA-49492.

We acknowledge the contribution of the Virginia Twin Registry, now part of the Mid-Atlantic Twin Registry (MATR), to ascertainment of participants for this study. The MATR, directed by J. Silberg and L. Eaves, has received support from the National Institutes of Health, the Carman Trust, and the W. M. Keck, John Templeton, and Robert Wood Johnson Foundations.

References

- Akaike H. Factor analysis and AIC. *Psychometrika* 1987;52:317–332.
- American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 3rd, rev. Washington, DC: Author; 1987.
- American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 4th. Washington, DC: Author; 1994.
- American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 4th, text rev. Washington, DC: Author; 2000.
- Cuthbert BN. Dimensional models of psychopathology: Research agenda and clinical utility. *Journal of Abnormal Psychology* 2005;114:565–569. [PubMed: 16351380]
- Frances A. Problems in defining clinical significance in epidemiological studies. *Archives of General Psychiatry* 1998;55:119. [PubMed: 9477923]
- Kendler, KS.; Prescott, CA. Genes, environment, and psychopathology: Understanding the causes of psychiatric and substance use disorders. New York: Guilford Press; 2006.
- Kessler, RC. National comorbidity survey, 1990–1992. 2nd ICPSR. Ann Arbor, MI: Interuniversity Consortium for Political and Social Research; 2002. Data file
- Kessler RC, McGonagle KA, Zhao S, Nelson CB, Hughes M, Eshleman S, Wittchen HU, et al. Lifetime and 12-month prevalence of *DSM-III-R* psychiatric disorders in the United States. *Archives of General Psychiatry* 1994;51:8–19. [PubMed: 8279933]
- Krueger RF, Hicks BM, Patrick CJ, Carlson SR, Iacono WG, McGue M. Etiological connections among substance dependence, antisocial behavior, and personality: Modeling the externalizing spectrum. *Journal of Abnormal Psychology* 2002;111:411–424. [PubMed: 12150417]
- Krueger RF, Markon KE, Patrick CJ, Iacono WG. Externalizing psychopathology in adulthood: A dimensional-spectrum conceptualization and its implications for *DSM-V*. *Journal of Abnormal Psychology* 2005;114:537–550. [PubMed: 16351376]
- Kubarych TS, Aggen SH, Hetttema JM, Kendler KS, Neale MC. Endorsement frequencies and factor structure of *DSM-III-R* and *DSM-IV* generalized anxiety disorder symptoms in women: Implications for future research, classification, clinical practice, and comorbidity. *International Journal of Methods in Psychiatric Research* 2005;14:69–81. [PubMed: 16175876]

- Lubke GH, Dolan CV, Neale MC. Implications of absence of measurement invariance for detecting sex limitation and genotype by environment interaction. *Twin Research* 2004;7:292–298. [PubMed: 15193174]
- MacCallum, RC. Model specification: Procedures, strategies, and related issues. In: Hoyle, RH., editor. *Structural equation modeling: Concepts, issues and applications*. Thousand Oaks, CA: Sage; 1995.
- Muthén BO. Psychometric evaluation of diagnostic criteria: Application to a two-dimensional model of alcohol abuse and dependence. *Drug and Alcohol Dependence* 1996;4:101–112.
- Neale MC, Aggen SH, Maes HH, Kubarych TS, Schmitt JE. Methodological issues in the assessment of substance use phenotypes. *Addictive Behaviors* 2006;31:1010–1034. [PubMed: 16723188]
- Neale, MC.; Boker, SM.; Xie, G.; Maes, HH. *Mx: Statistical modeling*. Richmond: Department of Psychiatry, Medical College of Virginia, Virginia Commonwealth University; 2002.
- Pickles A, Angold A. Natural categories or fundamental dimensions: On carving nature at the joints and the rearticulation of psychopathology. *Development and Psychopathology* 2003;15:529–551. [PubMed: 14582931]
- Rao, DC.; Gu, C. Principles and methods in the study of complex phenotypes. In: Benjamin, J.; Ebstein, RP.; Belmaker, RH., editors. *Molecular genetics and the human personality*. Washington, DC: American Psychiatric Publishing; 2002.
- Rebollo I, de Moor MHM, Dolan CV, Boomsma DI. Phenotypic factor analysis of family data: Correction of the bias due to dependency. *Twin Research and Human Genetics* 2006;9:367–376. [PubMed: 16790147]
- Regier DA, Kaelber CT, Rae DS, Farmer ME, Knauper B, Kessler RC, et al. Limitations of diagnostic criteria assessment instruments for mental disorders: Implications for research and policy. *Archives of General Psychiatry* 1998;55:109–115. [PubMed: 9477922]
- Rodebaugh TL, Woods CM, Heimberg RG, Liebowitz MR, Schneier FR. The factor structure and screening utility of the Social Interaction Anxiety Scale. *Psychological Assessment* 2006;18:231–237. [PubMed: 16768601]
- Spitzer, RL.; Williams, JBW. *Structured Clinical Interview for DSM–III–R (SCID)*. New York: Biometrics Research Department, New York Psychiatric Institute; 1985.
- Takane T, DeLeeuw J. On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika* 1987;52:393–408.
- Wirth RJ, Edwards MC. Item factor analysis: Current approaches and future directions. *Psychological Methods* 2007;12:58–79. [PubMed: 17402812]
- World Health Organization. *Composite International Diagnostic Instrument*. Geneva, Switzerland: World Health Organization; 1993.

Table 1
Demographic Characteristics of VATSPSUD and NCS Samples

Characteristic	VATSPSUD	NCS
Total <i>N</i>	2163	4263 ^a
Age		
<i>M</i>	30.1	33.3
<i>SD</i>	7.6	10.6
Years of education		
<i>M</i>	13.5	12.7
<i>SD</i>	2.1	2.3
Ethnicity (%)		
White	100	75.1
African American		12.5
Hispanic		9.1
Other		3.3
Marital status (%)		
Married	67	49.0
Cohabiting	5	5.5
Separated or divorced	10	16.8
Widowed	>1	1
Single	17	27.7
Religious affiliation (%)		
Protestant	83	55
Catholic	9	28
Jewish	1	
Other	5	7
None	2	10

Note. VATSPSUD = Virginia Adult Twin Study of Psychiatric and Substance Use Disorders. NCS = National Comorbidity Study.

^aRepresents women and girls only.

Table 2
Model Comparison of Effect of Age, Study, and Age $\times$ Study Interaction on Liability DSM-IV Generalized Anxiety Disorder (Probes Only)

Model number	Model	-2LnL	df	$\Delta\chi^2$	Δdf	AIC
1	Full measurement invariance (baseline)	8,095.67	8,000			-7,904.32
Effects on latent mean						
2	Age (vs. Model 1)	8,084.15	7,999	11.52**	1	-7,913.85
3	Study (vs. Model 1)	8,038.02	7,999	57.65***	1	-7,959.98
4	Age and study (vs. Model 1)	8,038.26	7,998	57.41***	2	-7,957.74
5	Age, Study, and Age $\times$ Study interaction (vs. Model 1)	8,033.17	7,997	62.51***	3	-7,960.83
Effects on thresholds						
6	Age (vs. Model 2)	8,043.33	7,994	40.82***	5	-7,944.67
7	Study (vs. Model 3)	7,971.81	7,994	66.21***	5	-8,016.19
8	Age and study (vs. Model 4)	7,948.74	7,998	89.52***	10	-8,027.26
9	Age, Study, and Age $\times$ Study interaction (vs. Model 5)	7,937.07	7,982	96.10***	15	-8,026.93
Effects on latent variance						
10	Age (vs. Model 1)	8,086.15	7,999	9.53*	1	-7,911.85
11	Study (vs. Model 1)	8,092.95	7,999	2.72	1	-7,905.05
12	Age and study (vs. Model 1)	8,081.53	7,998	14.14**	2	-7,914.47
13	Age, Study, and Age $\times$ Study interaction (vs. Model 1)	8,081.54	7,997	14.13**	3	-7,912.46
Effects on factor loadings						
14	Age (vs. Model 10)	8,082.45	7,994	3.69	5	-7,905.54
15	Study (vs. Model 11)	8,079.74	7,994	13.21*	5	-7,908.26
16	Age and study (vs. Model 12)	8,070.65	7,998	10.88	10	-7,905.35

Model number	Model	-2LnL	df	$\Delta\chi^2$	Δdf	AIC
17	Age, Study, and Age $\times$ Study interaction (vs. Model 13)	8,066.74	7,982	14.80	15	-7,897.26

Note. N = 6,426. The most parsimonious models are shown in boldface. *DSM-IV Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition*; AIC = Akaike information criterion.

* $p < .05$.
** $p < .01$.
*** $p < .001$.

Table 3
Model Comparison of Effect of Age, Study, and Age $\times$ Study Interaction on Liability to DSM–IV Generalized Anxiety Disorder (Including Stems)

Model number	Model	–2LnL	df	$\Delta\chi^2$	Δdf	AIC
1	Full measurement invariance (baseline)	14,247.28	14,408			–14,568.72
Effects on latent mean						
2	Age (vs. Model 1)	14,246.62	14,407	0.65	1	–14,567.38
3	Study (vs. Model 1)	14,244.44	14,407	2.84	1	–14,569.56
4	Age and study (vs. Model 1)	14,243.31	14,406	3.96	2	–14,568.68
5	Age, Study, and Age $\times$ Study interaction (vs. Model 1)	14,224.12	14,405	23.15***	3	–14,585.88
Effects on thresholds						
6	Age (vs. Model 2)	14,207.82	14,402	38.80***	5	–14,596.18
7	Study (vs. Model 3)	14,178.99	14,402	65.45***	5	–14,625.01
8	Age and study (vs. Model 4)	14,156.31	14,396	87.00***	10	14,635.69
9	Age, Study, and Age $\times$ Study interaction (vs. Model 5)	14,146.46	14,390	77.67***	15	–14,633.54
Effects on latent variance						
10	Age (vs. Model 1)	14,246.04	14,407	1.24	1	–14,567.96
11	Study (vs. Model 1)	14,241.95	14,407	5.32*	1	–14,572.04
12	Age and study (vs. Model 1)	14,241.73	14,406	5.55	2	–14,570.27
13	Age, Study, and Age $\times$ Study interaction (vs. Model 1)	14,237.74	14,405	9.53*	3	–14,572.26
Effects on factor loadings						
14	Age (vs. Model 10)	14,219.07	14,402	26.95***	5	–14,584.93
15	Study (vs. Model 11)	14,196.12	14,402	45.74***	5	–14,607.88
16	Age and study (vs. Model 12)	14,179.55	14,396	62.18***	10	–14,612.45

Model number	Model	-2LnL	df	$\Delta\chi^2$	Δdf	AIC
17	Age, Study, and Age $\times$ Study interaction (vs. Model 13)	14,171.84	14,390	65.90***	15	-14,608.16

Note. N = 6,426. The most parsimonious models are shown in boldface. DSM-IV = *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition*; AIC = Akaike information criterion.

* $p < .05$.

*** $p < .001$.

Table 4
DSM–IV Generalized Anxiety Disorder Item-by-Item Tests for Effects of Age on Thresholds

Model	-2LnL	$\Delta\chi^2$	Δdf	AIC	Age effect
Age on latent mean (baseline)	14,246.62			-14,567.38	
Item free					
Did your muscles often feel tense, sore or achy?	14,246.34	0.28	1	-14,565.66	.08
Did you often tire easily?	14,226.54	20.08***	1	-14,585.45	-.74
Were you so nervous you had trouble concentrating?	14,246.18	0.44	1	-14,565.82	-.10
Did you often have trouble falling asleep?	14,245.93	0.69	1	-14,566.07	-.13
Were you often irritable or especially impatient?	14,227.42	19.21***	1	-14,584.58	.73
Did you feel restless or keyed up and on edge?	14,246.50	0.12	1	-14,565.50	-.09

Note. $N = 6,426$. The most parsimonious model is shown in boldface. *DSM–IV = Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition*; AIC = Akaike information criterion.

$p < .001$.

Table 5
DSM-IV Generalized Anxiety Disorder Item-by-Item Tests for Effects of Study on Thresholds

Model	-2LnL	$\Delta\chi^2$	Δdf	AIC	Study effect
Study on latent mean (baseline)	14,244.44			-14,569.56	
Item free					
Did your muscles often feel tense, sore or achy?	14,235.31	9.12**	1	-14,576.69	.20
Did you often tire easily?	14,239.15	5.29*	1	-14,572.85	-.16
Were you so nervous you had trouble concentrating?	14,225.04	19.40***	1	-14,586.96	-.31
Did you often have trouble falling asleep?	14,238.60	5.84*	1	-14,573.40	-.17
Were you often irritable or especially impatient?	14,213.21	31.22***	1	-14,598.79	.40
Did you feel restless or keyed up and on edge?	14,235.86	8.57**	1	-14,576.14	.28

Note. $N = 6,426$. The most parsimonious model is shown in boldface. *DSM-IV* = *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition*; AIC = Akaike information criterion.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

Table 6
DSM-IV Generalized Anxiety Disorder Item-by-Item Tests for Effects of Age and Study Together on Thresholds

Model	-2LnL	$\Delta\chi^2$	Δdf	AIC	Age effect	Study effect
Age and study on latent mean (baseline)	14,243.31			-14,568.68		
Item free						
Did your muscles often feel tense, sore or achy?	14,234.02	9.29*	2	-14,573.97	-.06	.22
Did you often tire easily?	14,223.82	19.49***	2	-14,584.18	-.69	-.05
Were you so nervous you had trouble concentrating?	14,222.31	21.00***	2	-14,585.69	.23	-.35
Did you often have trouble falling asleep?	14,237.39	5.92	2	-14,570.61	.05	-.18
Were you often irritable or especially impatient?	14,204.62	38.70***	2	-14,603.38	.49	.34
Did you feel restless or keyed up and on edge?	14,232.75	10.56**	2	-14,575.25	-.33	.33

Note. $N = 6,426$. The most parsimonious model is shown in boldface. *DSM-IV* = *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition*; AIC = Akaike information criterion.

*
 $p < .05$.

**
 $p < .01$.

 $p < .001$.

DSM-IV Generalized Anxiety Disorder Item-by-Item Tests for Effects of Age, Study, and Age $\times$ Study on Item Thresholds

Table 7

Model	-2LnL	$\Delta\chi^2$	Δdf	AIC	Age effect	Study effect	Interaction effect
Age, study, and interaction on latent variance (baseline)	14,224.06			-14,585.94			
Item free							
Did your muscles often feel tense, sore, or achy?	14,211.36	12.70**	3	-14,592.64	0.35	.53	-.70
Did you often tire easily?	14,202.61	21.45***	3	-14,601.38	-1.11	-.33	.65
Were you so nervous you had trouble concentrating?	14,200.97	23.09***	3	-14,603.03	0.48	-.23	-.35
Did you often have trouble falling asleep?	14,218.11	5.95	3	-14,585.88	0.02	-.23	.09
Were you often irritable or especially impatient?	14,185.10	38.96***	3	-14,618.90	0.55	.38	-.07
Did you feel restless or keyed up and on edge?	14,213.96	10.10*	3	-14,590.04	-0.38	.33	.07

Note. $N = 6,426$. The most parsimonious model is shown in boldface. *DSM-IV* = *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition*; AIC = Akaike information criterion.

*
 $p < .05$.

**
 $p < .01$.

 $p < .001$.

Table 8
DSM-IV Generalized Anxiety Disorder Item-by-Item Tests for Effects of Age on Factor Loadings

Model	-2LnL	$\Delta\chi^2$	Δdf	AIC	Age effect
Age on latent variance (baseline)	14,246.04			-14,567.96	
Item free					
Did your muscles often feel tense, sore or achy?	14,245.99	0.05	1	-14,566.01	.03
Did you often tire easily?	14,235.09	10.95**	1	-14,576.91	.38
Were you so nervous you had trouble concentrating?	14,245.52	0.52	1	-14,566.48	.08
Did you often have trouble falling asleep?	14,246.02	0.02	1	-14,565.98	.01
Were you often irritable or especially impatient?	14,227.90	18.13***	1	-14,584.10	-.47
Did you feel restless or keyed up and on edge?	14,245.97	0.07	1	-14,566.03	.03

Note. $N = 6,426$. The most parsimonious model is shown in boldface. AIC = Akaike information criterion.

**
 $p < .01$.

 $p < .001$.

Table 9
DSM-IV Generalized Anxiety Disorder Item-by-Item Tests for Effects of Study on Factor Loadings

Model	-2LnL	$\Delta\chi^2$	Δdf	AIC	Study effect
Study on latent variance (baseline)	14,241.95			-14,572.04	
Item free					
Did your muscles often feel tense, sore or achy?	14,238.55	3.32	1	-14,573.45	-0.10
Did you often tire easily?	14,233.45	8.42**	1	-14,578.55	0.16
Were you so nervous you had trouble concentrating?	14,229.69	12.18***	1	-14,582.31	0.17
Did you often have trouble falling asleep?	14,240.13	1.74	1	-14,571.87	0.07
Were you often irritable or especially impatient?	14,224.63	17.24***	1	-14,587.37	-0.21
Did you feel restless or keyed up and on edge?	14,237.78	4.09*	1	-14,574.22	-0.11

Note. N = 6,426. The most parsimonious model is shown in boldface. DSM-IV = Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition; AIC = Akaike information criterion.

* $p < .05$.
** $p < .01$.
*** $p < .001$.

Table 10
DSM-IV Generalized Anxiety Disorder Item-by-Item Tests for Effects of Age and Study Together on Factor Loadings

Model	-2LnL	$\Delta\chi^2$	Δdf	AIC	Age effect	Study effect
Age and study on latent variance (baseline)	14,241.73			-14,570.27		
Item free						
Did your muscles often feel tense, sore or achy?	14,237.83	3.90	2	-14,570.16	.12	-.12
Did you often tire easily?	14,229.23	12.50**	2	-14,578.77	.28	.11
Were you so nervous you had trouble concentrating?	14,228.01	13.72**	2	-14,579.99	-.10	.17
Did you often have trouble falling asleep?	14,239.79	1.94	2	-14,568.21	-.06	.08
Were you often irritable or especially impatient?	14,215.61	26.12***	2	-14,592.39	-.37	-.15
Did you feel restless or keyed up and on edge?	14,235.02	6.71*	2	-14,572.98	.26	-.16

Note. $N = 6,426$. The most parsimonious model is shown in boldface. *DSM-IV* = *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition*; AIC = Akaike information criterion.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

Table 11
DSM-IV Generalized Anxiety Disorder Item-by-Item Tests for Effects of Age, Study, and Age $\times$ Study on Factor Loadings

Model	-2LnL	$\Delta\chi^2$	Δdf	AIC	Age effect	Study effect	Interaction effect
Age, study, and interaction on latent variance (baseline)	14,237.74			-14,572.26			
Item free							
Did your muscles often feel tense, sore or achy?	14,233.63	4.11	3	-14,570.37	-.03	-.16	.14
Did you often tire easily?	14,222.83	14.92**	3	-14,581.17	.57	.23	-.32
Were you so nervous you had trouble concentrating?	14,221.83	15.91**	3	-14,582.17	-.20	.08	.16
Did you often have trouble falling asleep?	14,235.36	2.38	3	-14,568.64	.13	.16	-.22
Were you often irritable or especially impatient?	14,212.33	25.40***	3	-14,591.66	-.35	-.15	.01
Did you feel restless or keyed up and on edge?	14,228.96	8.78*	3	-14,575.04	.24	-.14	.01

Note. $N = 6,426$. The most parsimonious model is shown in boldface. *DSM-IV = Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition*; AIC = Akaike information criterion.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

Table 12
Cross Tabulation of Diagnoses Based on Factor Scores Corrected for Measurement Noninvariance Versus Sum Scores: Full Sample

DSM-IV diagnosis (sum score ≥3)						
Factor score diagnosis	No		Yes		Total	
	n	%	n	%	n	%
No	5,189		51		5,240	81.54
Yes	52		1,134		1,186	18.46
Total	5,241	81.56	1,185	18.44	6,426	100.00
Full sample						
No	1,573		44		1,617	74.76
Yes	45		501		546	25.24
Total	1,618	74.80	545	25.20	2,163	100
NCS						
No	3,607		0		3,607	84.61
Yes	14		640		656	15.39
Total	3,623	84.99	640	15.01	4,263	100

Note. DSM-IV = *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition*; VATSPSUD = Virginia Adult Twin Study of Psychiatric and Substance Use Disorders; NCS = National Comorbidity Survey.