www.AUSTRALIANACADEMICPRESS.com.au

ARTICLE AVAILABLE ONLINE
Twin Research and Human Genetics
Volume 14 ▮ Number 6 ▮ pp. 524–538

Differential Age and Sex Effects in the Assessment of Major Depression: A Population-Based Twin Item Analysis of the DSM Criteria

Steven H. Aggen, Kenneth S. Kendler, Thomas S. Kubarych and Michael C. Neale
Virginia Institute for Psychiatric and Behavioral Genetics, Medical College of Virginia, Virginia Commonwealth University, Richmond, United States of America

A twin item factor analytic model was developed to test for the presence of noninvariant age, sex, and age by sex interaction effects on the individual DSM-III-R criteria for major depression (MD). Based on 1-year reports, six of the nine MD criteria and duration requirement were found to have covariate factor loading and/or threshold effects that significantly deviated from their corresponding factor level expectations. A significant age effect was found for the binary duration variable factor loading. The 'loss of interest', 'weight problems' and 'psychomotor problems' criteria all displayed forms of threshold only effects. 'Depressed mood', 'fatigue', and 'feeling worthless' had more complex patterns that included both factor loading and threshold effects. A significant factor age by sex interaction effect indicating an increasing female mean difference with age was found to be largely associated with the presence of differential threshold covariate effects. Disagreement between estimated factor scores and DSM-derived affected vs. unaffected classification was ~ 1.3%. Status on the duration requirement was found to be the one feature common to all discrepancies. The MD criteria set provided maximum information for calibrating MD factor scores in the scale region where discrepancies occurred. The dimensional modeling results are discussed in the broader context of epidemiological research and clinical assessment of major depression.

■ **Keywords:** item analysis, twin model, measurement invariance, DSM-III-R MD criteria, factor scores

One of the most robust findings in psychiatric epidemiology is the higher diagnostic prevalence of major depression (MD) in women compared to men (Bebbington et al., 1998; Nolen-Hoeksema, 1990; Piccinelli & Wilkinson, 2000; Weissman & Klerman, 1977). However, the interpretation of such a comparison depends on whether the designated diagnostic criteria equivalently assess MD in the two sexes (Rutter et al., 2003). Cross-sectional comparisons can be further complicated if the MD criteria are differentially influenced by age. If the criteria do not maintain a consistent and coherent relationship to the MD phenotype in the context of these and other relevant covariates, it is possible that different MD clinical features may give rise to diagnostic classifications, thereby increasing phenotypic heterogeneity across both individuals and studies.

Much of the evidence for MD differences comes from comparisons of rates of subjects meeting full diagnostic criteria (Blazer et al., 1994; Kendler et al., 1993; Nolen-Hoeksema, 1987). For MD, the criteria specified in the Diagnostic and Statistical Manual (DSM) (American Psychiatric Association, 1987; American Psychiatric Association, 1994) is a common choice in both clinical and research settings (Beauchaine, 2003). The classification procedure in effect 'counts up' all positive criteria in the set. Aggregating binary criteria in such a way imply rather strong assumptions about how the criteria relate to the disorder phenotype (Neale et al., 2005). First, since each criterion is given a unit weight, all criteria are in effect treated as equally representative of the disorder when determining if the diagnostic threshold is satisfied. Second, it assumes a consistency of both the inter- and intra-criteria properties across sampling and other

RECEIVED 05 November, 2010; ACCEPTED 07 October, 2011.

ADDRESS FOR CORRESPONDENCE: Steven H. Aggen PhD, Department of Psychiatry, Virginia Commonwealth University, PO Box 980126, Richmond, VA 23298-0126. E-mail: saggen@vcu.edu

possible selection conditions. These assumptions seem unlikely, a priori, given the face content of the different MD criteria and their possible differential expression across subpopulations. Nevertheless, in data analysis using the fully syndromal classification as the outcome, these assumptions are often implicitly assumed to hold but are rarely examined.

Within a classification system, it is difficult to evaluate these assumptions in a rigorous manner. However, if a latent variable modeling perspective is considered (Bollen, 2002; Borsboom et al., 2003), a set of testable hypotheses are available to evaluate both the construct validity (McArdle & Prescott, 1992) and measurement equivalence (ME) (Drasgow & Kanfer, 1985) of the individual MD criteria characteristics. This requires a conceptual shift in how the criterion level information is viewed and utilized. Rather than serving as unit weighted counts to establish a mutually exclusive dichotomous affected vs. unaffected classification, the MD disorder phenotype is conceived of as ordered individual differences on an unobserved continuous risk variable. If the observed relationships among the criteria can be predominately accounted for by a unidimensional factor structure (McDonald, 1981), the estimated criteria characteristics can be used to calibrate individual differences on the latent variable. It is these individual criterion characteristics that provide the basis for examining how each criterion 'functions' and relates to the MD disorder phenotype.

In latent variable modeling, ME can be evaluated through a series of restrictive *measurement invariance* (MI) hypotheses (Horn & McArdle, 1992; Meredith & Teresi, 2006). Failures of MI indicate the presence of conditional dependence of the item characteristics (e.g., item 'discrimination' and 'difficulty' parameters to be described later) on group membership (Millsap & Everson, 1993). A scale is said to be nonequivalent across subpopulations if persons with the same factor scores have different expected observed scale scores (Mellenbergh, 1989). Thus, it follows that persons in different subpopulations have not been scaled in the 'same way' on the latent variable. If measurement equivalence cannot be demonstrated, person and group differences may be confounded with other features that are not relevant to defining the construct. Conversely, if measurement equivalence does hold, there is a much stronger basis for making valid inference based on group difference that are found (Thissen et al., 1986).

Investigating MI statistically is typically done in one of two ways. One is through tests of factorial invariance using the common factor model (Meredith, 1993) and a second is by way of differential item functioning using item response theory (IRT) modeling (Holland & Wainer, 1993). The two approaches have much in common (Muthen & Asparouhov, 2002; Raju et al., 2002; Reise et al., 1993). In the case of dichotomous variables, the common factor model is formally equivalent to a 2-parameter normal ogive IRT model (Takane & de Leeuw, 1987). Both approaches estimate the item characteristic parameters that calibrate uniform individual person differences on the theoretical construct the items were intended to assess. For a collection of items to be considered as indicators of a single hypothetical construct, they should predominately fit a unidimensional structure (Hattie, 1984). Unidimensionality is typically defined by the principle of *local independence* (McDonald & Mok, 1995). It states that after accounting for variation in the latent variable (factor), the individual items should be statistically independent within narrow adjacent regions along the factor scale.

Two item parameters are of primary interest. The first is the item discrimination parameter (typically denoted as a or α in IRT models) and the second is the item difficulty (labeled b or β) parameter. Their counterparts in confirmatory factor analysis (CFA) are the factor loading (λ) and item threshold parameters (τ) respectively (Wirth & Edwards, 2007). The latter model and terminology is used here. Factor loadings give the increment of change on each of the items for a one unit linear change on the latent construct (factor). Larger values indicate stronger relationships. The threshold parameter is defined as the point on the latent scale where an item provides maximal information to distinguish scores above and below this location. The threshold coincides with the inflection point — the location on the factor scale where an item (criterion) has a 50% chance of being endorsed. Factor loadings and thresholds are jointly considered when investigating measurement invariance.

Present Study

A single-group item factor modeling approach (Neale et al., 2006) is used to investigate the MI properties of the nine DSM-II-R MD criteria and duration requirement with respect to age, sex and their interaction. Comparisons based on two conceptual distinctions in the model are used to test for the presence of noninvariant effects. Model fits with estimated covariate effects on the factor variance and mean are compared with models allowing separate covariate effects for each MD criteria factor loading and threshold respectively. Significant improvements in fit for the second set of models over the first suggest failures of MI. Additional testing is carried out to isolate the sources of the differential age, sex, and age by sex interaction effects. Discrepancies between algorithmic derived DSM-III-R affected vs. unaffected classifications and estimated factor scores are also examined.

Methods

Sample

Data for these item analyses come from two related studies of (1) female–female (FF), and (2) male–male and male–female (MMMF) twins from the Virginia Twin Registry

https://doi.org/10.1375/twin.14.6.524 Published online by Cambridge University Press

(Kendler & Prescott, 1999). The Virginia Twin Registry is a population-based register formed from a systematic review of all birth certificates in the Commonwealth of Virginia from 1918 onwards. Twins were eligible for participation in each of the studies if one or both twins were successfully matched to birth records and were born between 1940 and 1974 and were Caucasian.

The FF interviews were conducted between 1988 and 1989. Of the same-sex female twins deemed eligible, 92% (N = 2,163) participated. Interviews were conducted face-to-face (90%) or by phone (10%). Age at time of interview ranged from 18 to 55 with a mean of 30.1 (±7.6). Of the N = 9,417 MMMF eligible twins contacted between 1993 and 1996, N = 6,812 (72.3%) completed the interview. Age ranged from 19 to 57 with a mean of 35.5 (±9.1). For the total sample (N = 8,975), N = 5,090 were males mean age 35.5 (±9.2) and N = 3,885 were females mean age 32.5 (±8.6). Lifetime prevalence for meeting DSM-IIII-R criteria for MD was 29.5% for males and 36.7% for females. MD prevalence for the one year time period examined in this study was 10.2% for males and 11.6% for females. Reliabilities for these samples have been reported elsewhere (Kendler & Prescott, 2006). After a full explanation of the research protocol, signed consent forms were obtained prior to all face-to-face interviews and verbal assent was obtained for all telephone interviews. The two members of a twin pair were interviewed by different interviewers who were blind to clinical information about the co-twin.

The number of complete and partial (singleton) twin pairs in the sample was N = 5,055. There were N = 1,462 complete monozygotic (MZ) twin pairs and N = 284 MZ singletons. For dizygotic (DZ) twins, there were N = 2,509 complete pairs and N = 800 DZ singletons.

DSM-III-R MD Criteria

Using an adaptation of the SCID interview (Spitzer & Williams, 1985), each participant was asked to report if they had experienced any of the 14 disaggregated DSM-III-R criteria A symptoms over the 12 months prior to the interview. No skip-outs were used so every respondent was asked to provide a response for every MD criteria. Responses were record as either presence (1) or absence (0).

The 14 disaggregated DSM-III-R MD criteria assessed in the interview were: (1) depressed mood, (2) markedly diminished interest, (3a) significant weight loss or (3b) weight gain or (3c) increased appetite or (3d) decreased appetite (weight problems), (4a) insomnia or (4b) hypersomnia (sleep problems) (5a) psychomotor agitation or (5b) psychomotor retardation (psychomotor problems); (6) fatigue–loss of energy, (7) feelings of worthlessness, (8) inability to concentrate, and (9) recurrent thoughts of death. An additional 10th binary item was created to code for the DSM-III-R duration criteria indicating whether the syndrome (temporal clustering of symptoms) persisted for a minimum of 14 days. Numbers followed by a letter indicate how some of the disaggregated criteria were combined to get the nine DSM-III-R criteria used to obtain a diagnosis and included in the analysis. The separate weight and appetite (3a-d), sleep (4a-b), and psychomotor (5a-b) criteria were respectively collapsed to form new binary variables. If any one of the criteria within a set was positive, the new variable was coded as being present.

Each MD criterion had to meet three requirements to be scored positive. First, each symptom criteria had to have been experienced in the year prior to the interview. Second, positively endorsed criteria must have occurred in temporal proximity of one another. This ensured that positive criteria clustered in time to form a syndrome. If only a single symptom was reported, it was retained for the item analyses reported here. Finally, positive criteria were excluded if they were reported to be associated with physical illnesses or the taking of medication. Table 1 presents a summary of the endorsement proportions for each of the nine DSM-III-R MD criteria and duration requirement for males (columns 2) and females (columns 4). The ratios of male to female endorsements are reported in column 3 in descending order.

Item Statistical Model

A path diagram of the twin model used to test for differential covariate effects on the nine DSM-III-R MD criteria A and duration requirement is presented in Figure 1. Observed variables are drawn as boxes (□), latent variables (factors) are solid circles (○), triangles (Δ) represent constants for estimating means, diamonds (◊) indicate definition variables (Neale et al., 2004) that allow observed covariates to moderate the individual criteria properties in the model. Single headed arrows (→) indicate linear regressions and double headed arrows (↔) represent variances or covariances.

TABLE 1

Endorsement Proportions for the Nine DSM-III-R MD Criteria A and Duration Requirement Reported by Males and Females Over a 1-Year Time Period

MD Symptom Criteria	Endorse Prop M	Endorse M/F	Endorse Prop F
2 Loss of interest	0.224	0.891	0.252
9 Thoughts of death	0.038	0.856	0.045
5 Psychomotor problems	0.189	0.828	0.228
8 Cannot concentrate	0.106	0.828	0.128
4 Sleep problems	0.201	0.820	0.245
10 Duration criterion	0.266	0.785	0.339
7 Feelings worthless	0.099	0.767	0.129
1 Depressed mood	0.299	0.742	0.404
6 Fatigue	0.185	0.739	0.250
3 Weight problems	0.176	0.656	0.269

Note: Proportions are for N = 5,090 males and N = 3,885 females. Symptom endorsements are arranged in descending order according to the male-female endorsement proportion ratio.

https://doi.org/10.1375/twin.14.6.524 Published online by Cambridge University Press

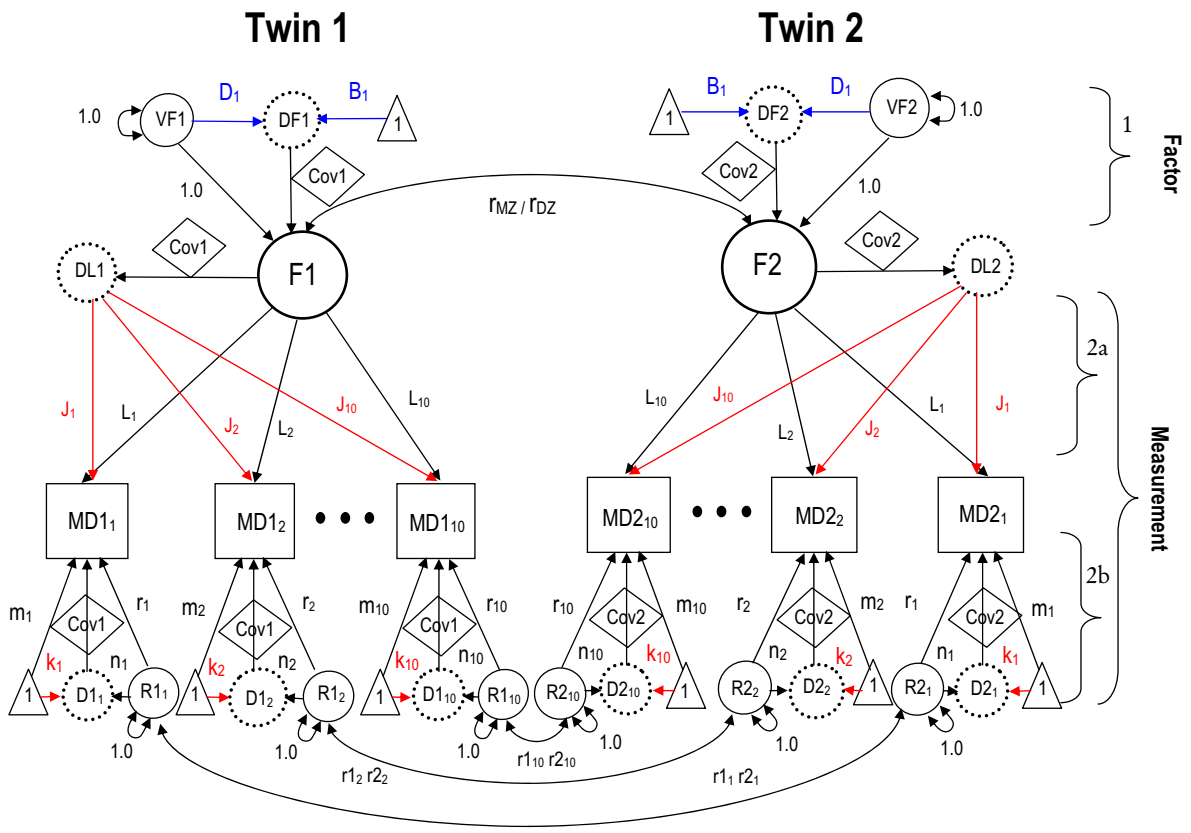

FIGURE 1

Twin model for testing differential effects of age, sex, and age by sex interaction on MD criteria factor loadings and thresholds.

As a conceptual aid, the diagram is divided into sections. The top portion of the model (1) shows how the estimation of the age, sex, and age by sex interaction covariate effects on the factor variance and mean are implemented. The B path estimates the covariate effects on the latent mean whereas the D path estimates the covariate effects for the factor variance. In this application, the factor variance is fixed to unity to satisfy the identification constraint necessary in all latent variable models. Fixing a single factor loading to unity can also be used to satisfy the identification constraint, but this parameterization can complicate attempts to isolate non-invariance features in the model (Cheung & Rensvold, 1999).

The two measurement components are labeled 2a and 2b. The first (2a) shows the factor loadings (L_i) for each MD criteria ($MD\#_x$) on the MD factor for each twin (F1 and F2). These are linear coefficients indicating the strength of relationship between each criterion and the MD factor. Factor loadings are analogous to the normal ogive 2-parameter IRT discrimination parameters and index how sharply each item discriminates individual differences on the factor. The single-headed arrows (J_i) from the latent nodes (DL#) to the factor loadings estimate the covariate effects. These are the linear effect sizes of how the covariate predictions of each factor loading deviate from the expected factor level covariate effects. Diamonds (Cov#) depict definition variables that incorporate the observed covariates into the model.

2b shows how the threshold structure is estimated. Each MD criterion has a single threshold (binary coding) location. Threshold estimates are the paths (m_i) pointing from the triangles ('1') to the observed MD criteria variables. Each threshold is allowed to have a different covariate effect through path k_i. As with the factor loadings, these effects quantify the direction and magnitude of the covariate effects on the individual MD criteria thresholds that are inconsistent with the single factor level covariate mean effects. To complete the measurement model, circles labeled $R\#_x$ with double headed arrows denote residual variances (i.e., specific plus random error variance).

The twin structural model (1) allows separate MZ and DZ correlations (r_{MZ}/r_{DZ}) between the twin1 and twin 2 common factors and (2) estimates twin1 and twin2 residual correlations for the same criterion (e.g., $r1_1$, $r2_1$). Parameter labels with subscripts (e.g., the B_1, D_1, L_i, J_i, and K_i) are constrained to be equal across the corresponding twin 1 and twin 2 parameters. Model parameters without subscripts can take on different values across twin members.

https://doi.org/10.1375/twin.14.6.524 Published online by Cambridge University Press

Each binary MD criterion is modeled as latent continuous response variable with a single threshold. Let y_i be the observed binary variables and y^*_i the corresponding latent continuous liability variables. The within person y^*_is (only the y_is are shown in Figure 1 as boxes labeled MD#$_{\#}$) can be expressed as:

$$y_i^* = \left(L_i + J_{ij}Cov_j\right)(F) + \left(R_i + K_{ij}Cov_j\right)$$

where F is the latent MD common liability factor, L_i are the $i = 1, 2, \ldots m$ factor loadings for the 9 MD and duration criteria, J_i are the specific covariate effects cov_j on each of the factor loadings (2a of Figure 1), and R_i are the $i = 1, 2, \ldots m$ residuals with specific K_{ij} effects for covariates cov_j. Across twin1 and twin 2, the inter-factor correlations

$$Cov(F_1,F_2) = \begin{cases} r_{MZ}: & if\ MZ\ twin\ pair \\ r_{DZ}: & if\ DZ\ twin\ pair \end{cases}$$

are separately estimated for MZ and DZ twins. MD twin factors F_t are assumed to be independent of the individual criteria residuals, both within and across twins:

$$Cov\left(F_t,R_{ij}\right) = 0$$

Twin 1 and twin 2 residuals for the same criteria may covary but no correlations are allowed for different criteria

$$Cov\left(R_{ij},R_{kl}\right) = \begin{cases} 0 & if\ i \neq k \\ r_iR_{MZ} & if\ i = k\ and\ MZ\ twin\ pair \\ r_iR_{DZ} & if\ i = k\ and\ DZ\ twin\ pair \end{cases}$$

Assuming a multivariate normal distribution underlies the expression of MD symptoms, a multiple threshold model can be used to model the observed pattern of item responses in each twin pair. The likelihood of a particular response pattern can be computed by integrating the multivariate normal distribution, and this likelihood can be maximized over the parameters of the model. Thus, MD criterion y_i is positive/endorsed (coded 1) if y^*_i is $\geq \tau^*$; otherwise 0, where τ is an estimated threshold. Accordingly, the likelihood for a pair of twins s with response vector (v_{i1}, v_{i2})

$$\int_{tp_{i1}}^{tq_{i1}} \cdots \int_{tp_{m2}}^{tq_{m2}} \varphi\left(y_{i1},y_{i2}\right)\ \ d_{m2} \ldots d_{11}$$

where (neglecting the subscript 1 or 2 denoting twin 1 or twin 2)

$$t_{qi} = \begin{cases} t_i & if\ v_i = 0 \\ +\infty & if\ v_i = 1 \end{cases}$$

and

$$t_{pi} = \begin{cases} -\infty & if\ v_i = 0 \\ t_i & if\ v_i = 1 \end{cases}$$

The multivariate normal density $\varphi(y_{i1}, y_{i2})$ depends on the item means and covariances, which are predicted by the structural equation model, and may be derived using matrix algebra or path tracing rules (Neale & Cardon, 1992). The covariate effects on the common factor mean and variance part of the model (Figure 1-1) can be expressed as

$$F_{\alpha\ j} = \alpha^* + B_jCov_j$$

$$F_{\sigma j} = \sigma^* + D_jCov_j$$

where $F_{\alpha j}$ and $F_{\sigma j}$ are the adjusted (for covariates measured on individual j) factor mean and variance, α^* and σ^* are the unadjusted factor mean and variance, and B_i and D_i are, respectively, vectors of linear regression parameters estimating the effects of the covariates on the factor means and covariances.

A series of nested comparisons based on the model shown in Figure 1 were formulated to test for differential age, sex, and age by sex interaction effects on the MD criteria. First, the fit for a baseline was established. This model represents the most parsimonious structure having no age, sex, or age by sex interaction effects at any level (i.e., factor or item). Second, model fits with age, sex, and age by sex interaction effects on the factor variance and mean are obtained. Because binary observed variables prohibit the estimation of factor loading and threshold covariate effects for all criteria simultaneously, separate tests were carried out for the factor variance vs. factor loadings and factor mean vs. thresholds respectively. To isolate the sources of differential factor loadings and thresholds covariate effects, additional model comparisons were performed.

The approach used here follows the methodology discussed in Neale et al. (2006) that has been used in several other recent research studies (Aggen et al., 2009; Kubarych et al., 2008; Kubarych et al., 2010). The Mx software (Neale et al., 2004) was use to implement a full-information (Bock et al., 1988) marginal maximum likelihood (Bock & Aitkin, 1981) procedure for estimating covariate effects for factor loadings and thresholds using raw data. The latent factor distribution was specified using a ten-point Gauss-Hermite quadrature. This approach has been shown to have certain advantages for hypotheses testing in latent variable models (Schmitt et al., 2006).

The modeling approach used here departs somewhat from conventional measurement invariance testing. No partitioning of the sample is done as is the case with multiple-group testing. Definition variables are used to estimate covariate effects using the full sample. This makes it possible to estimate effects with continuous variables (e.g., age) without having to impose arbitrary cut-offs (e.g., young vs. old). Definition variables also facilitate the simultaneous estimation of a number of covariates within the model. In the present study, the model developed makes it straightforward to include age, sex and their interaction as moderators of both the factor and criterion level characteristics.

https://doi.org/10.1375/twin.14.6.524 Published online by Cambridge University Press

Results

Unidimensionality

An important prerequisite for treating the MD criteria as indicators of a common disorder construct is that a single dominant factor adequately accounts for the pattern of observed associations among the criteria. The robust weighted least squares mean and variance adjusted estimator for categorical data in the Mplus software (Muthen & Muthen, 2004) was used to test for a unidimensional structure. For both the female (CFI = .99, TLI = .99, RMSEA = 0.04) and male (CFI = .99, TLI = .97, RMSEA = 0.03) (Bentler, 1990; McDonald & Marsh, 1990) samples, omnibus fit indices supported the unidimensional hypothesis (Aggen et al., 2005). Factor loadings ranged from 0.70 for 'thoughts of death' in females to 0.93 for 'depressed mood' in males. CFAs were also fit for the twin pair data constraining factor loadings to be invariant across twin 1 and twin 2. Results were nearly identical to those for the model fit to the individual record data.

Measurement Invariance Tests

Figure 2 is a graphical display of the likelihood space showing the relative differences in fit for all nested model comparisons examined. Model-data misfit is expressed as negative twice the log likelihood (-2lnL) and plotted on the Y-axis. The number of free (estimated) parameters is given on the X-axis. The upper graphic displays changes in misfit for models with age, sex, and age by sex interaction effects on the factor variance versus models allowing covariate effects on the individual MD criteria factor loadings. The lower graphic shows these same model comparison fits for the factor mean and criteria thresholds.

Although this graphical form of presenting model fitting results departs from the more familiar tabular format, we see several advantages to this type of presentation. First, relative differences in model-data misfit for all nested model comparisons can be visually comprehended as a gestalt. The amount of change in the -2lnL is expressed by the steepness of the line segments connecting the -2lnL values for models with more parameters. Steeper lines indicate greater improvement in fit per the additional parameters estimated.

Model misfits (-2lnL) used as comparison benchmarks are shown as circled numbers. Thin dotted lines originating from these reference models are contours of equivalent Akaike's Information Criterion (AIC) (Akaike, 1981; Akaike, 1987). These AIC contours provide an alternative index for evaluating models by balancing overall fit and model complexity (i.e., number of parameters estimated). Lines with double lettered labels represent likelihood-ratio chi-square difference tests. If certain regularity conditions are met, differences in -2lnL for appropriately nested models are distributed asymptotically as chi-squared. Solid lines indicate significant likelihood-ratio difference tests ($p < .05$) whereas dashed-dotted lines denote non-significant tests.

The baseline model (circled '1' labeled '1-No Moderation') has 20 estimated parameters (10 factor loadings and 10 thresholds). This model produced a -2lnL of 63538.4 and, although being the most parsimonious model, it produced the worst fit to the data. In the upper portion of Figure 2 (Loadings), the green labeled lines 'AV', 'BV', and 'CV' denote reductions in model misfit for models sequentially estimating age, sex, and an age by sex interaction on the factor variance. The single effect of age (AV) on the factor variance significantly improved the fit ($\Delta\chi^2_{(1)} = 56.7$, $p = .000$). Adding a sex (BV; $\Delta\chi^2_{(1)} = 16.4$, $p = .000$) and an age by sex interaction factor variance effect (CV; $\Delta\chi^2_{(1)} = 6.1$, $p = .01$) were also significant.

The 'DL', 'EL', and 'FL' lines give the improvements in fit for models allowing covariates to directly moderate the factor loadings. The multivariate test of age moderating all MD criteria factor loadings produced a significant improvement over a single age effect on the factor variance (DL; $\Delta\chi^2_{(9)} = 22.5$, $p = 0.007$). This was the only multivariate test for factor loadings to fall below the AIC contour line. Including all three covariate effects on the individual loadings (FL) did produce a significant likelihood ratio difference test (solid line) but did not fall below the corresponding AIC contour.

The lower portion of Figure 2 (Thresholds) shows model comparison results for criteria thresholds and factor mean covariate models. A different pattern of results is seen. A modest but significant effect of age on the factor mean (AM; $\Delta\chi^2_{(1)} = 9.7$, $p = .002$) was found. Adding a sex effect produced a substantial improvement in fit (BM; $\Delta\chi^2_{(1)} = 57.1$, $p = .000$); The age by sex interaction was also significant (CM; $\Delta\chi^2_{(1)} = 20.2$, $p = 0.000$). The multivariate covariate threshold tests were more pronounced and pervasive. All threshold moderation models (DT, ET, and FT) fall well below their corresponding AIC contours. Overall, the best fitting model by the AIC criterion was model 8 which allowed moderation of all criteria thresholds by all three covariates (FT; $\Delta\chi^2_{(27)} = 119.4$, $p = .000$).

Test results to determine which of the individual covariate effects were responsible for the multivariate findings are presented in Table 2. To be conservative, a model including all three covariate effects on the factor variance and factor mean (FV) and (FM) was used as the reference model. Factor variance and mean covariate effects were estimate in two ways: (1) age and sex effects were estimated separately; and (2) age, sex and an age by sex interaction effects were estimated jointly. AIC was used to determine improvement in model fit.

The age alone effect on the factor variance was significantly different from zero (0.10, [.08; .13]) indicating an increase in MD factor variance with age. Including age, sex, and age by sex interaction in the model, all had significant effects on the factor variance. Four MD criteria

https://doi.org/10.1375/twin.14.6.524 Published online by Cambridge University Press

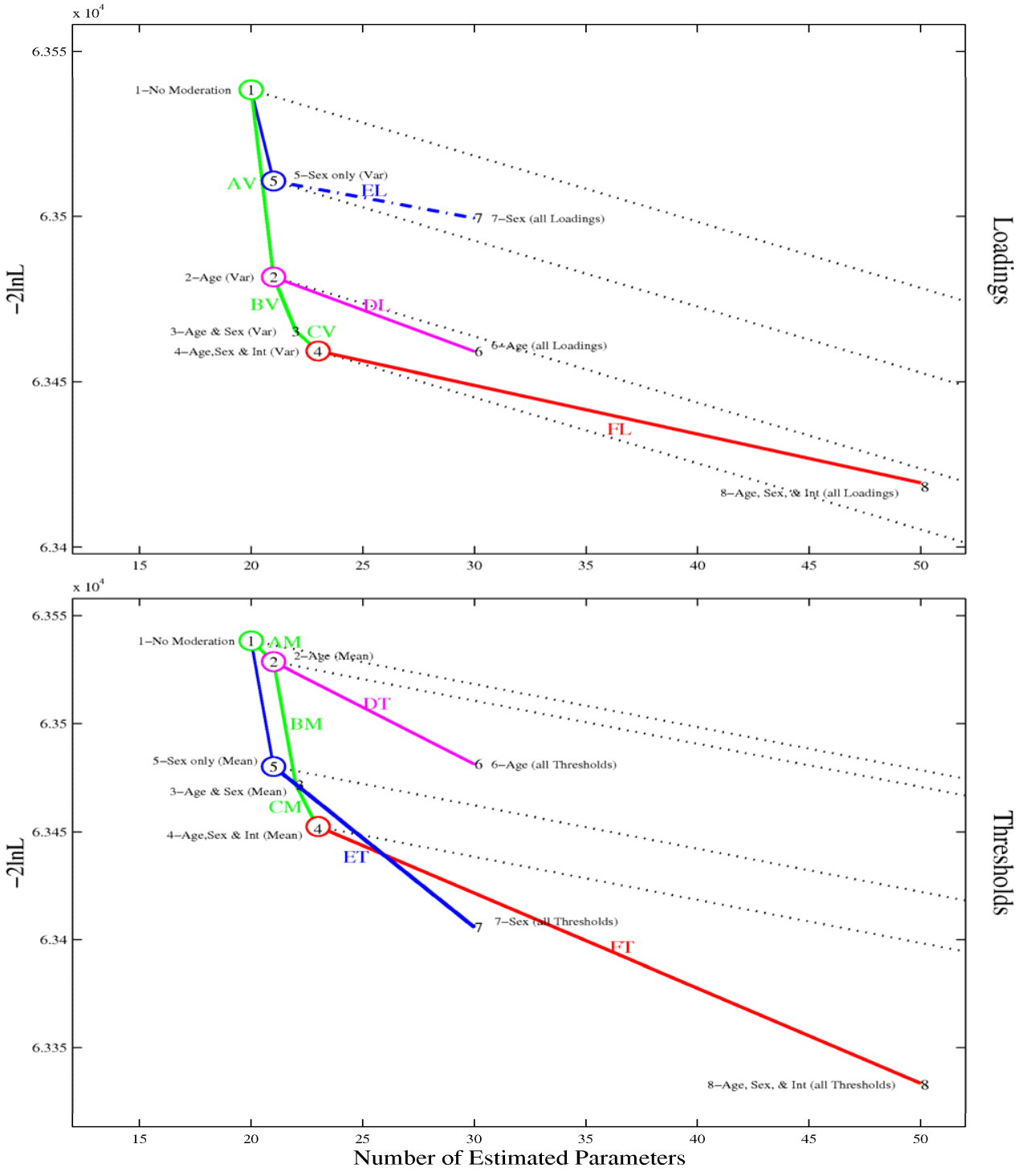

FIGURE 2

Plot of changes in the likelihood space for nested model comparisons. Negative twice the log likelihood (-2lnL) is plotted on the y-axis against number of parameters estimated on the x-axis. Letters lines denote improvements in fit for model comparisons. Models 2, 3, and 4 refer to age, sex, and age by sex interaction effects on the factor variance (upper) and the factor mean (lower). Letters A, B, and C (green) denote model comparisons for the factor variance. Circled numbers indicate baseline factor level covariate effect models (2, 4, and 5) and are connected to their counterpart factor loading (A) or threshold (B) moderation models. These changes in fit are labeled D (age, purple), E (sex, blue), and F (age by sex interaction, red). For example, model 2 allows age to affect the factor variance (A) or mean (B), whereas model 6 allows age to affect each factor loading (A) or criteria threshold (B). The likelihood ratio difference (connecting line D) shows the improvement in fit of model 6 over model 2 for the nine additional parameters (x-axis). The thin broken lines originating at each circled baseline model misfit show equal AIC 'contours' for corresponding additional parameters.

https://doi.org/10.1375/twin.14.6.524 Published online by Cambridge University Press

TABLE 2

Results of Testing for Effects of Age, Sex and Age x Sex Interaction on Each of Ten DSM-III-R MD Symptom Criteria

MD Symptom Criteria	-2lnL	*df*	$\Delta\chi^2$	Comparison Model	Ddf	$p\ \Delta\chi^2$	AIC	Age effect	Sex effect	Int. effect
(BL) Baseline Model (no moderation)	63538.4	90219	—	—	—	—	-116899.6	—	—	—
(FV) Covariate Effects Factor Variance	63459.3	90216	79.1	(BL)	3	**0.00**	-116972.7	0.071	-0.057	0.064
1. Depressed mood	63481.9	90213	6.0	(FV)	3	0.11	**-116972.7**	0.05	—	0.04
2. Loss of interest	63458.4	90213	0.9	(FV)	3	0.82	-116967.7	—	—	—
3. Weight problems	63458.6	90213	0.7	(FV)	3	0.87	-116967.4	—	—	—
4. Sleep problems	63453.9	90213	5.4	(FV)	3	0.15	-116972.1	—	—	—
5. Psychomotor problems	63456.4	90213	2.9	(FV)	3	0.41	-116969.6	—	—	—
6. Fatigue	63451.1	90213	8.2	(FV)	3	**0.04**	**-116974.9**	0.11	—	—
7. Feelings of worthlessness	63452.4	90213	6.9	(FV)	3	0.08	**-116973.6**	—	-0.04	-0.09
8. Cannot concentrate	63458.4	90213	0.9	(FV)	3	0.82	-116967.6	—	—	—
9. Thoughts of death	63453.9	90213	5.4	(FV)	3	0.15	-116972.1	—	—	—
10. Duration criteria (14 days or more)	63451.6	90213	7.7	(FV)	3	**0.05**	**-116974.4**	-0.07	—	—
(FM) Covariate effects factor mean	63452.4	90216	85.9	(BL)	3	**0.00**	-116979.6	-0.344	0.004	0.380
1. Depressed mood	63444.4	90213	8.0	(FM)	3	**0.05**	**-116981.6**	—	-0.05	—
2. Loss of interest	63423.2	90213	29.3	(FM)	3	**0.00**	**-117002.9**	—	0.12	0.21
3. Weight problems	63413.3	90213	39.1	(FM)	3	**0.00**	**-117012.7**	0.15	-0.15	-0.22
4. Sleep problems	63448.9	90213	3.4	(FM)	3	0.33	-116977.0	—	—	—
5. Psychomotor problems	63430.5	90213	21.9	(FM)	3	**0.00**	**-116995.5**	-0.26	—	—
6. Fatigue	63435.4	90213	17.0	(FM)	3	**0.00**	**-116990.6**	-0.17	-0.06	-0.17
7. Feelings of worthlessness	63436.5	90213	15.9	(FM)	3	**0.00**	**-116989.5**	0.23	—	—
8. Cannot concentrate	63446.9	90213	5.4	(FM)	3	0.15	-116979.0	—	—	—
9. Thoughts of death	63450.5	90213	1.9	(FM)	3	0.59	-116975.5	—	—	—
10. Duration criteria (14 days or more)	63450.8	90213	1.6	(FM)	3	0.66	-116975.2	—	—	—

Note: The upper panel shows differential effects on the factor loadings; the lower shows the effects on the thresholds. Estimated effect sizes are shown in the last three columns for significant effects only. -2lnL = Negative twice the log likelihood, *df* = degrees of freedom, $\Delta\chi^2$ = chi-square difference, Δdf = difference in degrees of freedom between compared models, $p\ \Delta\chi^2$ = probability associate with chi-square difference, AIC = Akaike's Information Criterion.

displayed some form of factor loading covariate moderation that departed from the factor variance covariate expectations ('depressed mood', 'fatigue', 'feelings of worthlessness', and the 'duration criteria').

For factor mean effects, sex by itself had a significant positive effect (0.15, [.11; .16]). Females, on average, have higher MD factor scores compared to males. However, this factor mean sex difference was noticeably reduced and rendered nonsignificant (0.01, [-.07; .09]) when the age by sex interaction was included. Six criteria ('depressed mood', 'loss of interest', 'weight problems', 'psychomotor problems', 'fatigue', and 'feelings of worthlessness') were found to have significant forms of differential threshold covariate moderation. These effects are given in the last three columns of Table 2 for factor loadings (upper) and thresholds (lower).

To further examine the patterns of differential age, sex, and interaction covariate effects for the MD criteria, bootstrapping was carried out. Using a twin model that included all significant factor mean and variance covariate effects plus all significant factor loading and threshold noninvariant effects, the model was refit five-hundred times to random samples drawn with replacement from the original data. Figure 3 displays these bootstrapping results. Noninvariant effects are expressed using 4 points with 95% confidence intervals (CI) for each MD criterion. Significant differential effects of factor loadings (A) and thresholds (B) are identified by the numbered points, (1) no covariate effects (males no age effects), (2) male age effect, (3) sex effect (female-male sex difference), and (4) age by sex interaction effect (sex effect plus the age by sex interaction effect).

The geometric shapes formed by the lines connecting the points provide a visual description of the nature of the differential covariate effects. Points without numbers and CIs that are identical indicate MD criteria with no significant differential effects (e.g., 'sleep problems'). An effect of age but not sex would appear as a parallelogram with horizontal red and blue broken lines. The 'fatigue' factor loading (A) and 'psychomotor' threshold (B) follow this pattern. Sex but no age effects appear as a parallelogram with offset horizontal green and purple lines (e.g., 'depressed mood' threshold (B)). Interaction effects can yield triangular or trapezoid shapes, such as is evident for the 'feeling worthless' factor loading (A) and 'weight problem' threshold (B). That factor loadings are, in general, estimated with less precision than are thresholds is evident from their wider confidence intervals.

https://doi.org/10.1375/twin.14.6.524 Published online by Cambridge University Press

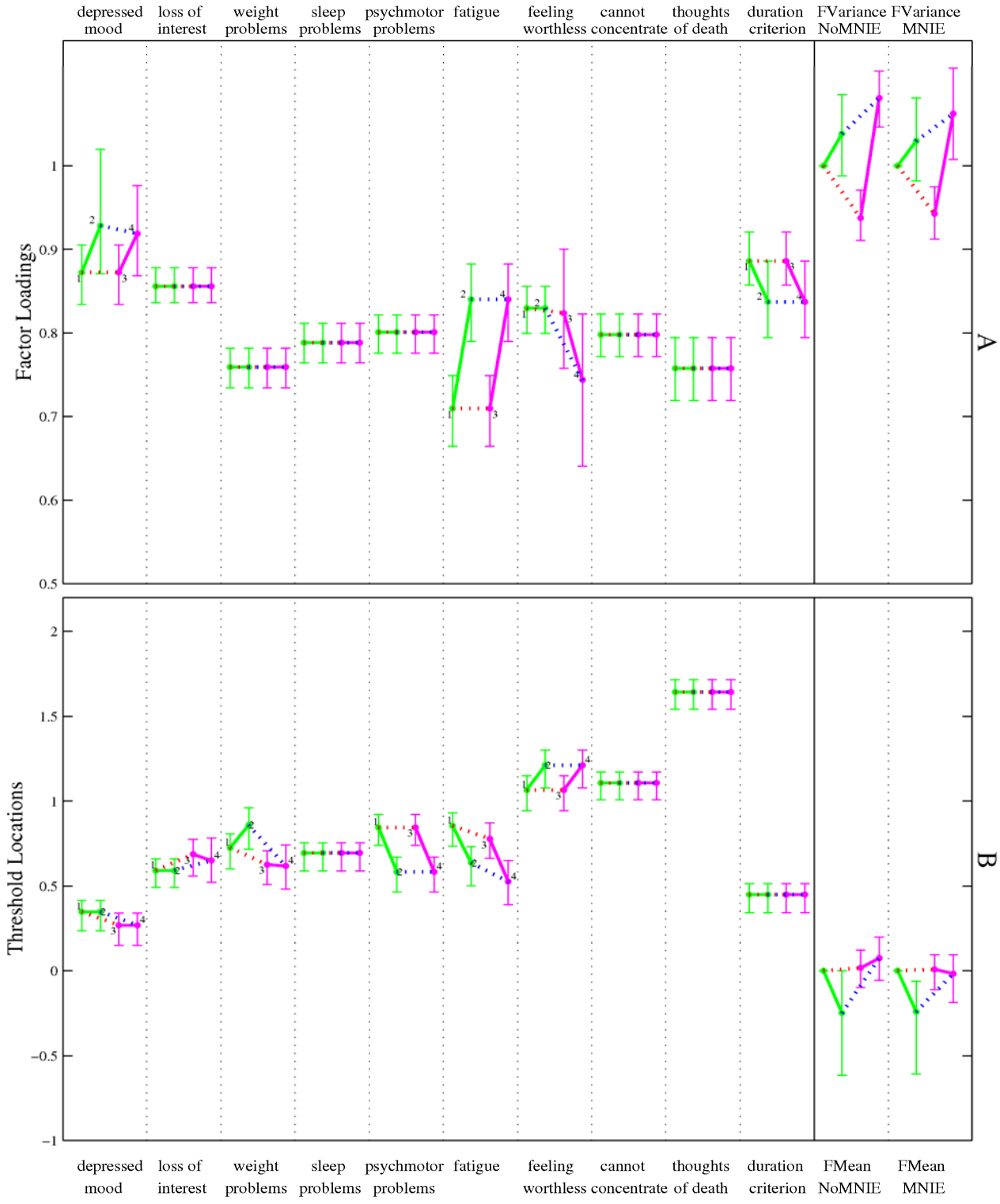

FIGURE 3

Plots of bootstrap estimates of factor loadings (upper A) and thresholds (lower B). MD symptom criterion identifiers are given at the top and bottom. For criteria with significant moderation, four labeled points give estimates and 95% CI's for 1) no covariate effects (males no age effects), (2) male age effect, 3) sex effect (female–male sex difference), and 4) age by sex interaction effect (sex effect plus the age by sex interaction effect). Red broken lines connect median bootstrap estimates for youngest males and females; broken blue lines connect the estimates for the oldest males and females plus the age by sex interaction. The geometric shapes formed by the lines describe the differential symptom functioning due to sex, age and their interaction.

https://doi.org/10.1375/twin.14.6.524 Published online by Cambridge University Press

All MD criteria factor loadings (Figure 3 A) had estimated values at or above 0.7. The 'depressed mood' and 'fatigue' criteria displayed unexpected increases in discrimination with age. For older individuals, these MD criteria discriminate individual differences on the MD factor more sharply than would be expected given the single MD factor variance effect of age. In contrast, the 'duration' item had an unexpected decrease in discrimination with age. Finally, the 'feeling worthless' criterion had a more complex moderation pattern that included an age by sex interaction effect (i.e., triangular shape). Compared to same aged males, females showed a significant decline in discriminating power for this criterion with increasing age.

Figure 2B shows unexpected covariate effect patterns for the MD criteria thresholds. Criteria thresholds were all located above the zero factor scale point or 'average' risk level. Therefore, all MD criteria predominately provided information about MD factor score differences towards the high end of the risk scale in this population-based sample. Several criteria had differential effect patterns suggesting the presence of an age by sex interactions (i.e., triangular shapes). The 'loss of interest', 'weight problems', and 'fatigue' criteria all followed such a pattern. A differential sex effect was found for 'depressed mood' with females tending to endorse this criterion more often (lower threshold) than expected compared to men when accounting for the factor sex mean effect. The 'psychomotor' threshold displayed a significant age effect — this criterion tended to be endorsed more by older twins than expected based on the single factor mean age effect. An age effect in the opposite direction was found for 'feeling worthless' with older twins reporting experiencing this MD symptom less often than expected.

The two columns to the far right show how covariate effects on the factor variance (Figure 3A) and mean (B) were impacted by the presence of measurement non-invariant effects at the criterion level. The pattern of covariate effects on the variance did not change but were somewhat reduced in magnitude. However, the age-by-sex interaction effect on the factor mean was noticeably altered suggesting a confounding with the differential threshold effects that were present. The likelihood ratio test comparing a model including all significant differential covariate effects against the baseline model produced a sizeable reduction in misfit ($\Delta\chi^2_{(17)} = 208.9$, $p < .001$).

MD factor scores were estimated under four different parameterizations: (1) the baseline model, (2) allowing factor mean and variance covariate effects, (3) allowing for only significant factor loading and threshold moderation effects, and (4) both 2 and 3. To evaluate the disagreement between factor scores and algorithmic derived affected/unaffected diagnostic classifications, the percentage of twins meeting DSM-III-R requirements for major depression (11.3%) was used as a cut-off for the corresponding highest rank ordered MD factor scores. Condition 1 produced disagreement for $N = 103$ twins, ~ 1.2% of the total sample. Under conditions 2, 3, and 4, discrepancies were $N = 132$ (~1.5%), $N = 112$ (~1.3%) and $N = 132$ (~1.5%) respectively. Adjusting for age, sex, and interaction effects on the factor and individual criteria increased the rate of disagreement. Disagreement for positive/negative diagnoses and factor scores below/above the 11.3% cutoff were found to be fairly symmetric.

Figure 4 is a histogram of estimated MD factor scores adjusted for significant non-invariance covariate effects. Factor scores are partitioned into the four possible categories obtained by crossing male/female with affected/unaffected status. Approximately one percent of the total sample ($N = 112$) had estimated factor scores that disagreed with the binary diagnoses. Discrepancies fell in the region marked by dashed vertical lines. The solid black line is the break point separating the two types of discrepancies. Below this line individuals had factor scores falling outside the top 11.3% of the distribution but were assigned an affected status by the diagnostic algorithm. Above this line discrepancies were reversed. From a measurement perspective, it of interest to note that the location separating the different types of discrepancies coincided with the maximum level of information provided by the 10 MD criteria. Factor scores in this region are the most precisely calibrated based on the criteria information. Also, all discrepancies had one feature in common – their status on the duration criterion. Cases with factor scores above the cut-off but classified as unaffected all failed to meet the 14-day minimum duration requirement. In contrast, cases classified as affected but their corresponding factor score fell below the cut-off, the duration criterion was met.

We note that when a set of items is essentially unidimensional (i.e., a single factor model fits adequately) as is the case here for the 9 DSM-III-R MD criteria and duration requirement, the estimated factor scores generally correlate highly with the more straightforward sum of the binary criteria that are used to determine diagnostic case status. However, when investigating the effects of covariates on MD criteria, one might find different item level effects for the common portion of the criteria (factor loadings) compared to those specific to each criterion. This would not be possible with the sum scores. Also, if the item set is multidimensional, sum scores can produce covariate effects that may be distorted and misleading because they ignore the structure present in the item associations. In other work we have established that maximum likelihood factor scores are more accurate estimates of true factor scores than are sum scores when responses are available from only a subset of the items, that is, there some items have missing data (Estabrook and Neale, submitted).

https://doi.org/10.1375/twin.14.6.524 Published online by Cambridge University Press

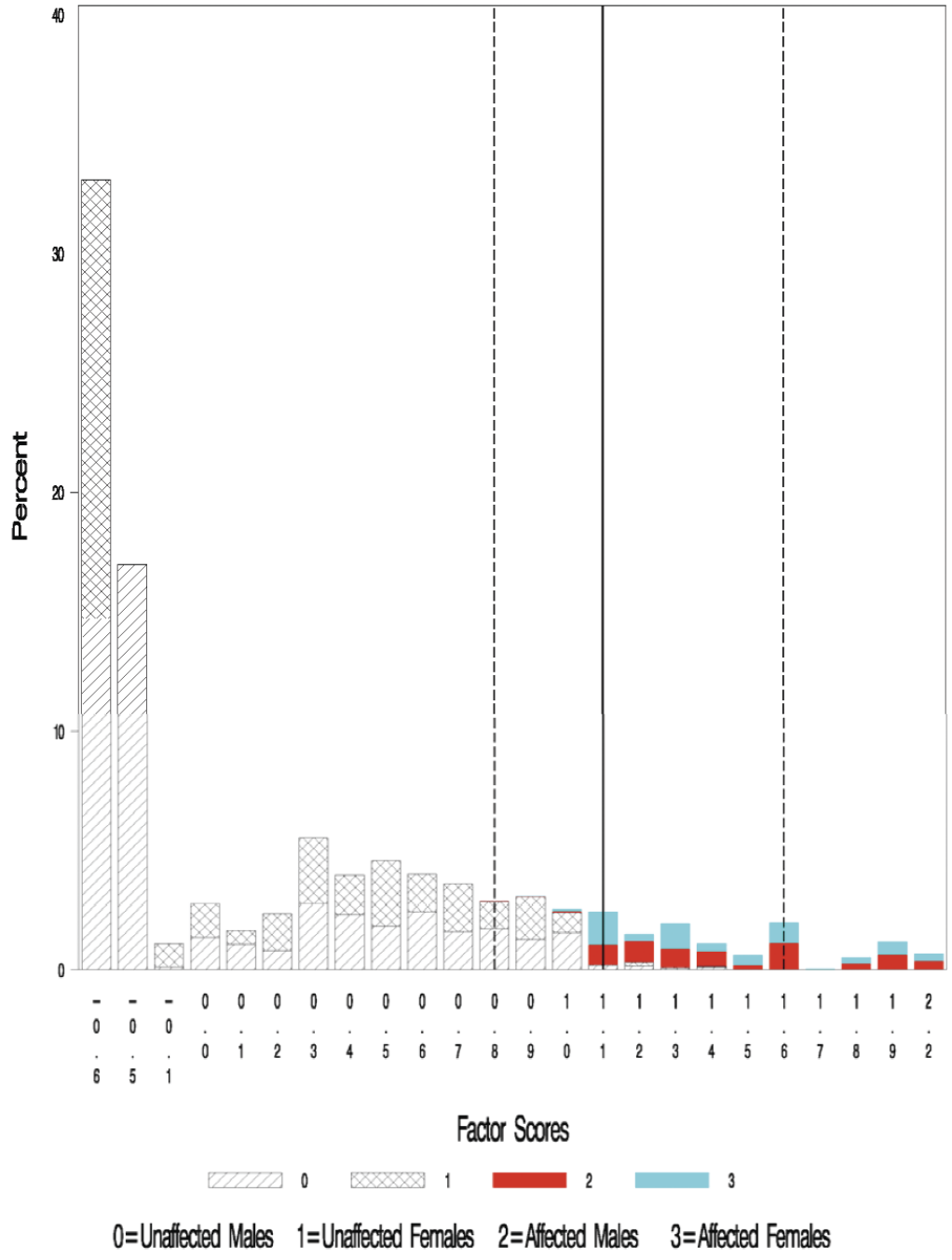

FIGURE 4

Histogram of discrepancies between estimated MD factor scores when considering both factor loading and threshold covariate noninvariant effects and DSM affected vs. unaffected classification effects.

Discussion

A single-group twin-item factor analytic modeling approach was used to investigate age, sex and age by sex interaction effects for the nine aggregated DSM-III-R MD criteria A and duration requirement. Models estimating covariate effects for each MD criterion factor loading and threshold were compared to models that included only covariate effects on the factor variance and mean respectively. Significant improvements in fit of the former models over the latter are evidence for the failure of measurement invariance. If multivariate tests were significant, additional tests were conducted to isolate the specific sources of the noninvariance.

Various forms of differential age, sex, and age by sex effects were found for six of the nine DSM-III-R criteria A and duration requirement. The 'fatigue' criteria factor

https://doi.org/10.1375/twin.14.6.524 Published online by Cambridge University Press

loading had a positive differential age effect. This criterion discriminated factor scores better for older twins than would be expected given the factor variance age covariate effects. The factor loading for 'depressed mood' displayed a significant age and interaction effect. The binary duration requirement (whether or not the syndrome persisted for 14 days) was found to have a negative differential age effect, such that discrimination was poorer for older twins. 'Feelings of worthlessness' had a more complicated pattern that included both sex and interaction effects. This criterion discriminated less than expected but only for older females.

Differential covariate effects were more apparent and pervasive for MD criteria thresholds. The 'depressed mood' criterion had a significant negative sex effect with females tending to report this clinical feature more often (lower threshold) than males given the factor sex mean effect. Since this criterion is one of the two designated hierarchical clinical features required for assigning a positive diagnosis, this finding suggests that females may have a disproportionately higher likelihood compared to males to meet the MD diagnostic threshold cut off. However, the form of differential moderation for 'loss of interest' (the other designated core feature) included both positive sex and age by sex interaction components. This pattern of noninvariant effects suggests higher endorsement rates for males compared to females but this difference declines with age. 'Psychomotor problems' and 'feelings of worthlessness' were differentially moderated by age in opposite directions with the former displaying greater than expect endorsement by older twins whereas the latter showed lower than expected endorsement rates (i.e., higher threshold). The 'weight' and 'fatigue' criteria displayed forms of noninvariant effects including all three covariates.

The combined impact of the significant differential covariate effects did not appear to have a noticeable impact on the pattern of covariate effects on the factor variance although effect sizes were attenuated when covariates were included for the factor loadings. However, impact on the factor mean effects was more evident. A significant age by sex interaction complicated a straightforward interpretation of the sex effect on the factor mean. A sex only effect on the factor mean was significant and consistent with the extensive literature on the fully syndromal condition. On average, compared to males, females were higher on the MD factor. However, this MD factor mean difference was reduced to a non-significant level when the age by sex interaction was included in the model. To further complicate the interpretation, the interaction term was found to be impacted by and possibly confound with the presence of significant differential threshold covariate effects. These rather specific and novel findings, although interesting and thought provoking, will require further study to determine if they are replicable. However, more generally, these findings do suggest a more complex relationship between the DSM-III-R symptom criteria and risk for (and the diagnosis of) MD. For example, the results reported here highlight a need to consider the designated clinical features of MD from a broader more developmentally oriented perspective. Although not longitudinal estimates, the linear age and sex by age interaction effects on the criterion characteristics suggest the presence of timing related characteristics that might be operative in the expression of the MD clinical features.

Examining disagreement between estimated MD factor scores and DSM algorithmic derived affected vs. unaffected classifications produced several noteworthy results. First, the discrepancy rate was ~1.3 percent. When interpreting this disagreement, it should be kept in mind that approximately 50% of this population-based twin sample reported no MD symptoms for the year prior to the interview. For this portion of the sample, agreement between the two methods is assured since all that are classified as unaffected had very low (but not identical due to the influence of covariate effects) estimated factor scores. Second, the number of discrepancies *increased* when factor scores were adjusted for all significant non-invariance effects. This is to be expected since the binary classification algorithm ignores and is insensitive to all these criteria level covariate effects. Third, all cases of disagreement had a common feature. If they were classified as affected but their factor scores were not in the top 11.3 % of the distribution, the 14 day minimum duration requirement was met. For the reverse case, if factor scores were above the cut-off but assigned an unaffected status, the duration requirement was not met. It was found that the region where discrepant cases were located on the factor scale coincided with the area of maximum information provided by the DSM-III-R criteria. That is, the DSM-III-R MD criteria calibrated MD factor scores with the most precision (i.e., least measurement error) in the region were discrepancies were observed.

These findings draw attention to several issues in the epidemiological assessment of major depression using the DSM-III-R diagnostic criteria. First, it is important to recognize that the individual criteria may be differentially sensitive to certain background and demographic features that depart from their overall collective power to delineate a single coherent MD disorder phenotype. These specific characteristics are as much a part of the MD disorder phenotype as are the common factor level effects. In fact, as highlighted in the analysis reported on here, the valid interpretation of the factor effects may depend on an awareness and understanding of any differential criteria level effects that may be present. Second, based on these dimensional measurement model results, the pervasive sex difference typically reported for diagnostic classification prevalence differences appears to be more complicated when examining effects at the criterion characteristic level. This highlights a key feature of conducting research and

https://doi.org/10.1375/twin.14.6.524 Published online by Cambridge University Press

statistical modeling. Results are dependent on what the model takes into account. Conditioning on important covariates that influence the target phenotype can alter the effects obtained and their interpretation.

Although the dimensional item level analytic models used here do not explicitly take into account the hierarchical and exclusionary conditions as is the case in the DSM classification system, there may be some distinct advantages to not doing so. For example, when including the 14-day minimum duration requirement as just another binary criterion for calibrating a person's location on the MD continuum, this binary variable turned out to be a perfect predictor of disagreement between estimated factor scores and binary MD classification. One way to view these results is that two new groups of people have been identified who suffer depression symptomatology. First there are those individuals with relatively mild but more persistent expression. Second there are those who experience more severe but rather short-lived affliction. Since the length of time a syndromal pattern of symptoms persists can be quite variable across persons, imposing a fixed cut-off may have undesirable and limiting consequences. Based on results reported here, imposing a fixed duration cut-off of 14 days obscured variation in patterns of MD symptoms and the individual differences they implied. Factor scores of those misclassified ranged between 0.8 and 1.6 units on the interval MD factor risk factor scale.

In psychometrics, the failure of measurement invariance is typically interpreted as an impediment to valid statistical inference — especially in the case of group comparisons. The presence of differential item functioning can confound and obscure 'true' population effects. However, failures of MI can be viewed from another perspective in the current context. Identifying how key covariates may differential moderate symptom expression is a first step in attempting to probe the more dynamic features that may be at work in the manifestation and course of the MD disorder phenotype. Understanding the nature of these criteria specific features can extend and inform substantive theory and interpretation. For example, criteria specific covariate effects that depart from the overall general effects may be of clinical interest. Also, by identifying and sorting out these more detailed criteria specific effects, another perspective is available for attempting to understand the developmental features of this complex psychiatric disorder.

Limitations

The twin data used here to examine the DSM-III-R MD symptom criteria for differential covariate effects were not collected using the typical hierarchical DSM skip-out structure. The presence or absence of all MD symptoms over the last year was asked of all twins. This strategy reduced the impact of missing data and selection effects typically present when hierarchical skip-outs are used but as a result includes information on probe criteria that would not be available when the fully syndromal condition is analyzed.

Although changes observed with age at interview may be the result of 'aging', the research design is cross-sectional. The correlation between birth-year and age is large, but not perfect in this study, so it is possible that some of the effects that we have attributed to age may be due to cohort effects. The male and female twins used in this study are from the Virginia Twin Registry — an all-Caucasian sample mainly localized to the state of Virginia. Thus, there may be limitations to how they generalize to other populations.

Finally, although we opted to investigate the aggregated 9 DSM-III-R symptom criteria for major depression because they are used to make diagnostic classifications in both clinical and research settings, it is important to acknowledge the potential impact of aggregation when modeling individual criteria. In preliminary dimensional analyses, the 14 disaggregated criteria were not unidimensional. Although the decreased weight and appetite criteria loaded substantially on the primary MD factor, the increased weight and appetite criteria did not and formed a second factor that included negative loadings (polarity) for the decreased weight and appetite criteria. Given the nature of these criteria, it seems possible differential sex and age effects could be present but obscured in the current analyses. Future research is needed to explore this possibility.

Acknowledgments

This work was supported by NIH grants MH-40828, MH-65322 and MH/AA/DA49492. We acknowledge the contribution of the Virginia Twin Registry, now part of the Mid-Atlantic Twin Registry (MATR), to ascertainment of subjects for this study. The MATR, directed by Dr. J. Silberg, has received support from the National Institutes of Health, the Carman Trust and the WM Keck, John Templeton and Robert Wood Johnson Foundations.

References

Aggen, S. H., Neale, M. C., & Kendler, K. S. (2005). DSM criteria for major depression: Evaluating symptom patterns using latent-trait item response models. *Psychological Medicine, 35*, 475–487.

Aggen, S. H., Neale, M. C., Røysamb, E., Reichborn-Kjennerud, T., & Kendler, K. S. (2009). A psychometric evaluation of the DSM-IV borderline personality disorder criteria: age and sex moderation of criterion functioning. *Psychological Medicine*, *39*, 1967–1978.

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika, 52*, 317–332.

Akaike, H. (1981). Likelihood of a model and information criteria. *Journal of Econometrics, 16*, 3–14.

American Psychiatric Association (1987). *Diagnostic and Statistical Manual of Mental Disorders, Revised Third Edition*. Washington, DC: American Psychiatric Association.

American Psychiatric Association (1994). *Diagnostic and Statistical Manual of Mental Disorders*, (4th ed.). Washington, DC: American Psychiatric Association.

Beauchaine, T. P. (2003). Taxometrics and developmental psychopathology. *Development and Psychopathology, 15,* 501–527.

Bebbington, P. E., Dunn, G., Jenkins, R., Lewis, G., Brugha, T. S., Farrall, M., & Meltzer, H. (1998). The influence of age and sex on the prevalence of depressive conditions: Report from the national survey of psychiatry morbidity. *Psychological Medicine, 28,* 9–19.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107,* 238–246.

Blazer, D. G., Kessler, R. C., McGonagle, K. A., & Swartz, M. S. (1994). The prevalence and distribution of major depression in a National Community Sample: The National Comorbidity Survey. *American Journal of Psychiatry, 151,* 979–986.

Bock, R. D. & Aitkin, M. (1981). Marginal maximum-likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46,* 443–459.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-Information item factor analysis. *Applied Psychological Measurement, 12,* 261–280.

Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology, 53,* 605–634.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review, 110,* 203–219.

Cheung, G. W. & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management, 25,* 1–27.

Drasgow, F. & Kanfer, R. (1985). Equivalence of psychological measurement in heterogeneous populations. *Journal of Applied Psychology, 70,* 662–680.

Estabrook, R. & Neale, M. C. (2011). A Comparison of Factor Score Estimation Methods in the Presence of Missing Data. (submitted).

Hattie, J. (1984). An empirical study of various indices for determining unidimensionality. *Multivariate Behavioral Research, 19,* 49–78.

Holland, P. W. & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.

Horn, J. L. & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18,* 117–144.

Kendler, K. S., Neale, M. C., Kessler, R. C., Heath, A. C., & Eaves, L. J. (1993). A longitudinal twin study of 1-year prevalence of major depression in women. *Archives of General Psychiatry, 50,* 843–852.

Kendler, K. S. & Prescott, C. A. (2006). *Genes, environment, and psychopathology: Understanding the causes of psychiatric and substance use disorders*. New York: Guilford Press.

Kendler, K. S. & Prescott, C. A. (1999). A population-based twin study of lifetime major depression in men and women. *Archives of General Psychiatry, 56,* 39–44.

Kubarych, T. S., Aggen, S. H., Hettema, J. M., Kendler, K. S., & Neale, M. C. (2008). Assessment of generalized anxiety disorder diagnostic criteria in the National Comorbidity Survey and Virginia Adult Twin Study of Psychiatric and Substance Use Disorders. *Psychological Assessment, 20,* 206–216.

Kubarych, T. S., Aggen, S. H., Kendler, K. S., Torgersen, S, Reichborn-Kjennerud, T., & Neale, M. C. (2010). Measurement non-invariance of DSM-IV narcissistic personality disorder criteria across age and sex in a population-based sample of Norwegian twins. *International Journal of Methods in Psychiatric Research. 19*, 156-166.

McArdle, J. J. & Prescott, C. A. (1992). Age-based construct-validation using structural equation modeling. *Experimental Aging Research, 18,* 87–115.

McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical & Statistical Psychology, 34,* 100–117.

McDonald, R. P. & Marsh, H. W. (1990). Choosing a multivariate model — noncentrality and goodness of fit. *Psychological Bulletin, 107,* 247–255.

McDonald, R. P. & Mok, M. M. C. (1995). Goodness-of-fit in item response models. *Multivariate Behavioral Research, 30,* 23–40.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research, 13,* 127–143.

Meredith, W. (1993). Measurement invariance, factor-analysis and factorial invariance. *Psychometrika, 58,* 525–543.

Meredith, W., & Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical Care, 44,* S69–S77.

Millsap, R. E., & Everson, H. T. (1993). Methodology review - statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17,* 297–334.

Muthen, B., & Asparouhov, T. (2002). Latent variable analysis with categorical outcomes: Multiple–group and growth modeling In Mplus. Mplus Web Notes: No.4 Version 5, December 9, 2002 [On-line].

Muthen, B. O., & Muthen, L. K. (2004). *Mplus user's guide* (3rd ed.). Los Angeles, CA: Muthen & Muthen.

Neale, M. C., & Cardon, L. R. (1992). Methodology for genetic studies of twins and families. Dordrecht, The Netherlands.

Neale, M. C., Aggen, S. H., Maes, H. H., Kubarych, T. S., & Schmitt, J. E. (2006). Methodological issues in the assessment of substance use phenotypes. *Addictive Behaviors, 31,* 1010–1034.

Neale, M. C., Boker, S. M., Xie, G., & Maes, H. H. (2004). *Mx: Statistical modeling* (6th ed.) Box 980126, Richmond, VA: Department of Psychiatry, Virginia Commonwealth University.

Neale, M. C., Lubke, G., Aggen, S. H., & Dolan, C. V. (2005). Problems with using sum scores for estimating variance

https://doi.org/10.1375/twin.14.6.524 Published online by Cambridge University Press

components: Contamination and measurement noninvariance. *Twin Research and Human Genetics, 8,* 553–568.

Nolen-Hoeksema, S. (1987). Sex differences in unipolar depression: Evidence and theory. *Psychological Bulletin, 101,* 259–282.

Nolen-Hoeksema, S. (1990). *Sex differences in depression*. Palo Alto, CA: Stanford University Press.

Piccinelli, M., & Wilkinson, G. (2000). Gender differences in depression — Critical review. *British Journal of Psychiatry, 177,* 486–492.

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 87,* 517–529.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor-analysis and item response theory — two approaches for exploring measurement invariance. *Psychological Bulletin, 114,* 552–566.

Rutter, M., Caspi, A., & Moffitt, T. E. (2003). Using sex differences in psychopathology to study causal mechanisms: Unifying issues and research strategies. *Journal of Child Psychology and Psychiatry, 44,* 1092–1115.

Schmitt, J. E., Mehta, P. D., Aggen, S. H., Kubarych, T. S., & Neale, M. C. (2006). Semi-nonparametric methods for detecting latent non-normality: A fusion of latent trait and ordered latent class modeling. *Multivariate Behavioral Research, 41,* 427–443.

Spitzer, R. L. & Williams, J. B. W. (1985). *Structured clinical interview for DSM-III-R (SCID)*. New York: Biometrics Research Department, New York State Psychiatric Institute.

Takane, Y. & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52,* 393–408.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin, 99,* 118–128.

Weissman, M. M., & Klerman, G. L. (1977). Sex differences and the epidemiology of depression. *Archives of General Psychiatry, 34,* 98–111.

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods, 12,* 58–79.

https://doi.org/10.1375/twin.14.6.524 Published online by Cambridge University Press