Psychological Medicine, 2005, **35**, 475–487. © 2004 Cambridge University Press
DOI: 10.1017/S0033291704003563 Printed in the United Kingdom

DSM criteria for major depression: evaluating symptom patterns using latent-trait item response models

STEVEN H. AGGEN*, MICHAEL C. NEALE AND KENNETH S. KENDLER

Virginia Institute for Psychiatric and Behavioral Genetics, Medical College of Virginia and Virginia Commonwealth University, Richmond, VA, USA

ABSTRACT

Background. Expert committees of clinicians have chosen diagnostic criteria for psychiatric disorders with little guidance from measurement theory or modern psychometric methods. The DSM-III-R criteria for major depression (MD) are examined to determine the degree to which latent trait item response models can extract additional useful information.

Method. The dimensionality and measurement properties of the 9 DSM-III-R criteria plus duration are evaluated using dichotomous factor analysis and the Rasch and 2 parameter logistic item response models. Quantitative liability scales are compared with a binary DSM-III-R diagnostic algorithm variable to determine the ramifications of using each approach.

Results. Factor and item response model results indicated the 10 MD criteria defined a reasonably coherent unidimensional scale of liability. However, person risk measurement was not optimal. Criteria thresholds were unevenly spaced leaving scale regions poorly measured. Criteria varied in discriminating levels of risk. Compared to a binary MD diagnosis, item response model (IRM) liability scales performed far better in (i) elucidating the relationship between MD symptoms and liability, (ii) predicting the personality trait of neuroticism and future depressive episodes and (iii) more precisely estimating heritability parameters.

Conclusions. Criteria for MD largely defined a single dimension of disease liability although the quality of person risk measurement was less clear. The quantitative item response scales were statistically superior in predicting relevant outcomes and estimating twin model parameters. Item response models that treat symptoms as ordered indicators of risk rather than as counts towards a diagnostic threshold more fully exploit the information available in symptom endorsement data patterns.

INTRODUCTION

Over the last 50 years, committees of expert clinicians have chosen the official diagnostic criteria for psychiatric disorders. More recently, the process has been complemented by empirical studies examining classical reliability and the validity of competing criteria sets (Kendler, 1990). However, quantitative measurement theory (Mitchell, 1997) and modern psychometrics (Embretson & Reise, 2000) have not played a prominent role in the criteria evaluation and selection process. The primary objective of the *Diagnostic Statistical Manual* (DSM; APA, 1987) criteria is to discriminate 'cases' from 'non-cases' rather than to define latent dimensions of liability.

In both clinical work and research, information is typically collected on the presence or absence of all symptom criteria for a given disorder but then collapsed into a single affected

* Address for correspondence: Dr Steven H. Aggen, Department of Psychiatry, PO Box 980126, Richmond, VA 23298-0126, USA.
(Email: saggen@hsc.vcu.edu)

versus unaffected classification. From a data analytical perspective, dichotomizing is usually inefficient (Cohen, 1983) and can produce misleading results (MacCallum *et al.* 2002). When collapsing multiple criteria into a single binary diagnosis, a large portion of the available information in the symptom endorsement patterns are not utilized (Wainwright *et al.* 1997). The decision to investigate psychiatric phenotypes using a categorical rather than dimensional orientation is best viewed as a choice since it has proven difficult to establish that a given disorder is more accurately represented as mutually exclusive classes or continuous variation (Pickles & Angold, 2003).

Although attempts have been made to bring attention to the potential benefits of applying latent trait methods in psychiatric and clinical settings (Duncan-Jones *et al.* 1986), there have been relatively few applications that examine the DSM criteria for major depression. Reiser (1989) discussed the use of item response models (IRMs) in the context of psychiatric epidemiology and applied a multivariate logistic regression approach to analyze the eight DSM-III B criteria for major depression. Muthén (1989*a*), using a dichotomous factor analytic approach, obtained a single factor solution for the DSM-III criteria for major depression (MD) from two sites of the ECA study. The weight/appetite and fatigue criteria were found to have the lowest loadings on the latent factor. Other applications of IRMs in psychiatry, particularly to self-report scales assessing symptoms of schizophrenia (e.g. Lewine *et al.* 1983; Bell *et al.* 1994), depression (Gibbons *et al.* 1993; Orlando *et al.* 2000), and, more generally, the relationship between symptoms and diagnoses (Grayson *et al.* 1987) are also noteworthy.

In this paper, latent trait item-response models (LT-IRMs; Gibbons *et al.* 1985) are applied to past year symptom data on DSM-III-R criteria for MD. The phrase item response models (IRMs) is used here instead of the more conventional item response theory (IRT) to denote an emphasis on models rather than theory (Goldstein & Wood, 1989). The analyses reported are motivated by two sets of questions. First, measurement concerns address how well the DSM-III-R criteria for MD stochastically work together. Do the criteria define one or more dimensions of MD liability? How informative are the criteria in reliably distinguishing individual differences in liability? Are criteria uniformly dispersed across the range of liability? This measurement perspective departs from how diagnostic classification algorithms utilize symptom data (Blashfield & Livesley, 1991).

A second set of questions concerns how much additional information can be obtained from LT-IRM constructed scales compared to a binary representation? In particular, how do any information gains relate to external predictor validation? We investigate this line of inquiry by jointly examining both representations of depression with respect to: (i) their linear regression on the personality trait of neuroticism (N) [which has been shown in many studies to be a key risk factor for MD (Angst & Clayton, 1986; Hirschfeld *et al.* 1989; Kendler *et al.* 1993)], (ii) their relative predictive power of future episodes of illness, and (iii) their differential precision in estimating heritability parameters in twin models.

METHOD

Sample and criteria

A sample of $n=2163$ Caucasian same-sex female twins from the Virginia Twin Registry is used. This 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 to participate if they were born between 1934 and 1971 and both members of the pair had previously responded to a mailed questionnaire (response rate $\sim 64\%$). Of the eligible twins, 91·9% were successfully interviewed. Of the completed interviews, 89·3% were completed face-to-face in the twins' home, with the remaining 10·7% (mostly twins living outside Virginia) interviewed by telephone. The mean age ($\pm$S.D.) of the sample at the time of interview was $30{\cdot}1 \pm 7{\cdot}6$ years and ranged from 17 to 55 years. Signed informed consent was obtained prior to all face-to-face interviews and verbal assent prior to all telephone interviews.

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 time

https://doi.org/10.1017/S0033291704003563 Published online by Cambridge University Press

of the interview. No skip-outs were used in this section of the interview so each respondent was asked about every symptom. Responses were recorded as binary indicating either the presence or absence of each symptom.

The 14 disaggregated DSM MD criteria are: (1) depressed mood, (2) markedly diminished interest, (3a) significant weight loss or (3b) significant weight gain or (3c) increased appetite or (3d) decreased appetite, (4a) insomnia or (4b) hypersomnia, (5a) psychomotor agitation or (5b) psychomotor retardation, (6) fatigue or loss of energy, (7) feelings of worthlessness, (8) inability to concentrate, and (9) recurrent thoughts of death. Numbers followed by a letter indicate how the 14 disaggregated symptoms were organized into the nine DSM-III-R diagnostic criteria for a MD episode. For these analyses, the individual weight and appetite (3a–d), sleep (4a, b), and psychomotor (5a, b) criteria were collapsed respectively within each grouping. If any one of the symptoms within the group was endorsed the criteria was marked as being present. This aggregation was performed to construct the IRM liability scales with the same information used to determine a DSM-III-R MD diagnosis.

Only item responses for diagnostic criteria meeting the following conditions were included in the analyses: (1) all positively endorsed symptoms must have occurred in temporal proximity to one another forming a syndrome cluster. If only a single symptom was positively endorsed, the symptom was retained for analysis. (2) Positive symptoms were not included if they were associated with physical illnesses or the taking of medication. To follow the DSM-III-R requirements as closely as possible, a tenth binary item was included indicating whether or not a syndrome persisted for a minimum of 2 weeks.

Measurement models

IRMs simultaneously calibrate items and measure persons on a common latent scale – here labeled the liability (or risk) to MD. The Rasch (Rasch, 1960, 1966) and 2-parameter logistic (2PL) models (Birnbaum, 1968), although developed under different measurement traditions, are used to examine the measurement characteristics of DSM-III-R criteria for MD. The Rasch model specifies the requirements observations must meet in order to construct additive linear scales (Andrich, 1989). If the Rasch model holds, both persons and criteria can be ordered on a common unit preserving scale. Diagnostics can be used to evaluate how closely criteria conform to these measurement requirements. Thus, the Rasch model is a special latent trait model evaluating the degree to which the DSM-III-R criteria for MD can construct a quantitative measure. The Winsteps software (Linacre & Wright, 2000) is used to estimate Rasch parameters and obtain misfit indices.

The second approach is more closely aligned with methods discussed under the rubric of IRT (Lord & Novick, 1969). Observations determine the relative plausibility and explanatory merit of alternative IRMs. Models are rejected or retained in accordance with how closely model expectations reproduce the structure implied by data. Global data-model misfit is typically assessed with single valued discrepancy indexes. In this respect, the 2PL model introduced by Birnbaum (1968) is used to fit a less restrictive model allowing discrimination parameters to vary. Estimating additional parameters to further account for data features differs from the Rasch objective in which observations must conform to the specification of the model in order to achieve linear measurement. Parameter estimates and model likelihood fits are obtained with the software Multilog (Thissen, 1991). Multilog uses an iterative expectation-maximization (EM) algorithm (Dempster *et al*. 1977) to implement the method of marginal maximum-likelihood estimation (Bock & Aitkin, 1981).

The two IRMs offer different interpretations of how the symptom criteria relate to person liability. Under the Rasch model, criteria must operate in a specific manner if the resulting scale is to be additive and linear. If the Rasch model requirements are satisfied, the observed summed score is a sufficient statistic for independently estimating criteria (thresholds only) and person parameters. Thus, there is some justification criteria can be 'counted' although the summed integer values, not to be confused with the Rasch scaled values, are not of equal interval. For Rasch scaled scores, the marginal symptom count (regardless of pattern) is transformed and numerically re-expressed such that meaningful comparisons can be made between persons.

Since symptom counting is a part of almost all operationalized hierarchical diagnostic systems when non-essential criteria are aggregated, Rasch results are directly relevant to determining whether the practice of counting symptoms is reasonable.

The 2PL model differs in several ways. At the cost of giving up the fundamental measurement objectives achieved under the Rasch approach, the 2PL model provides additional criteria information by allowing discrimination parameters to differ. The liability scale is thus segmented in a more refined manner. Each unique endorsement pattern now can potentially yield a separate person liability score. However, in order to obtain parameter estimates for this model requires that a distributional form be specified for the latent liability.

Applications of IRMs to symptom endorsement data depart from their more familiar use in the cognitive ability domain. While the conventional IRT terminology of item difficulty clearly applies to the proportion of individuals able to answer a test item correctly, it is not appropriate when applied to the percentage of individuals in a population endorsing a psychiatric symptom. Symptoms are not 'difficult' in the sense that a test item is in which a certain level of ability is needed to arrive at the correct answer (maximum performance assessment). Symptom endorsements typically reflect retrospective personal recall of whether or not a particular symptom was experienced or not (self-referenced recall assessment). The nature of the response processes underlying these two data- gathering tasks is quite different. Therefore, we adopt the term 'liability threshold' to describe the estimated point on the scale of disease liability where a symptom has a 50% probability of being endorsed.

To assess the dimensionality of the 10 MD criteria, exploratory and confirmatory dichotomous factor analyses as implemented in the software package Mplus (Muthén & Muthén, 2001) are performed. The common factor model applied to dichotomous data (Christoffersson, 1975; Bartholomew, 1980) decomposes the matrix of tetrachoric correlations into common and specific/error latent variables.

A second and potentially more informative approach to assessing dimensionality along with measurement quality is to evaluate the consistency between the observed endorsement patterns given the Rasch model criteria calibrations. The infit and outfit indices (Wright & Stone, 1979; Linacre & Wright, 2000) are designed to identify criteria that poorly differentiate MD risk due to response patterns that depart from model expectations. Infit is an information-weighted summary statistic sensitive to departures from expectations for persons with estimated liability scores close to a criterion's estimated threshold. Outfit detects unexpected endorsements by persons whose risk scores are far away from a criterion's threshold. Both are expressed as mean-square (χ^2's divided by their degrees of freedom) statistics with an expectation of unity and a range of 0 to $+\infty$. Values less than 1 indicate criteria that are stochastically more consistent with the model than expected and values greater than 1 indicate excess noise. A mean-square fit statistic (ZSTD) is standardized to approximate a theoretical mean 0 variance 1 distribution (Wright & Stone, 1979). Suggested mean square infit and outfit cut-off values for survey ratings and clinical observations are between 0·5 and 1·7. For more precise testing, cut-offs between 0·8 and 1·2 are recommended (Wright & Linacre, 1994).

Validation

In the absence of gold standards for assessing accuracy (Faraone & Tsuang, 1994), external validity is an important source of evidence to evaluate psychiatric constructs (Robins & Guze, 1970). To examine external validity, the two IRM-derived liability scales are compared with a binary diagnostic variable in three ways: predicting (1) Neuroticism scores obtained from a prior wave of data collection analyzed by linear regression and (2) a diagnosis of past year MD obtained in a follow-up interview analyzed by logistic regression. Statistical predictive power, effect sizes, and tests of statistical significance are examined in regression models that sequentially include age, the binary MD diagnosis, the IRM risk scales, and an interaction term. A third form of external validation examines the performance of the binary and quantitative liability variables in a twin modeling application. Mx (Neale *et al.* 1999) confidence intervals (CIs) are used to examine the precision of twin correlations and parameter estimates for additive genetic effects (A), shared

https://doi.org/10.1017/S0033291704003563 Published online by Cambridge University Press

Table 1. *Dichotomous common factor and Rasch model results testing the dimensionality and item performance for 10 DSM-III-R MD criteria*

	Confirmatory factor analyses		Rasch Out misfit statistics				
				Infit		Outfit	
	F1	Res	Thres.	MSQ	ZSTD	MSQ	ZSTD
1 Depressed	0·87	0·25	−1·80	0·99	−0·50	0·88	−1·85
2 Nointerest	0·84	0·30	−0·37	0·91	−3·23	0·88	−2·77
3 Weightapp	0·69	0·52	−0·58	1·14	4·80	1·19	4·23
4 Sleepprob	0·79	0·38	−0·21	0·95	−1·80	0·93	−1·65
5 Psycmotor	0·77	0·41	−0·18	0·97	−0·95	0·96	−0·90
6 Fatigue	0·71	0·50	−0·21	1·08	2·38	1·11	2·21
7 Worthless	0·75	0·45	0·90	0·99	−0·12	1·01	0·10
8 Concentra	0·74	0·45	0·95	0·99	−0·16	0·97	−0·41
9 Suicidal	0·72	0·49	2·47	1·06	0·62	0·93	−0·39
10 Duration	0·78	0·40	−0·96	0·98	−0·67	0·93	−1·57
χ^2	119, $p<0·000$						
df	35						
CFI	0·97						
TLI	0·96						
RMSEA	0·03						

The confirmatory factor solution was obtained in Mplus using a weighted least squares estimation method producing a robust χ^2 fit statistic. F1, factor loading point estimates for each criteria; Res, residual (unique & error) variance; Thres, Rasch threshold estimates obtained from Winsteps. Winsteps Rasch infit and outfit statistics: MSQ, mean square; ZSTD, mean square normalized to mean and variance expectations 0 and 1 respectively. CFI, Comparative Fit index; TLI, Tucker–Lewis index; RMSEA, root mean square error of approximations.

environmental factors (C) and individual-specific environmental effects plus error (E).

Factor solutions are evaluated by three fit-indices: the Tucker–Lewis Index (TLI; Marsh *et al.* 1988), the Comparative Fit Index (CFI; Bentler, 1990), and the root mean square error of approximation (RMSEA; Steiger, 1990). For the TLI and CFI, values between 0·90 and 0·95 are considered reasonable and values $>0·95$ as good. For the RMSEA, good fit approximations are $\leqslant 0·05$, while values $>0·10$ are considered poor (Browne & Cudeck, 1993).

RESULTS

Dimensionality and criteria functioning

Table 1 shows results for dichotomous confirmatory factor (left) and Rasch model (right) analyses. An initial exploratory factoring provided a reasonable one-factor solution (RMSEA = 0·03). The confirmatory factor results listed in Table 1 corroborate this result showing that a single common factor model satisfactorily accounts for the pattern of tetrachoric correlations among the criteria (CFI = 097, TLI = 0·96). The factor loadings are in general high ranging from 0·69 to 0·87. The weight/appetite and fatigue criteria had the lowest loadings suggesting these two criteria are less central to the core interpretation of the MD liability construct than are the other criteria.

The right portion of Table 1 gives Rasch threshold estimates and corresponding infit and outfit statistics. Of the 10 criteria, the Rasch infit and outfit statistics were positive and statistically significant for the weight/appetite (infit = 1·14, ZSTD = 4·80; outfit = 1·19, ZSTD = 4·23) and fatigue (infit = 1·08, ZSTD = 2·38; outfit = 1·11, ZSTD = 2·21) criteria. The significant negative infit and outfit statistics for the no-interest criteria indicate less stochastic variation than expected under the model. However, even for these criteria, the infit and outfit values fall within an acceptable range when using a $\pm 0·2$ (i.e. 0·8–1·2) cut-off. These Rasch item misfit statistics not only support a unidimensional interpretation of the DSM-III-R MD diagnostic criteria but also indicate the criteria function reasonably well in constructing an additive linear MD risk scale.

Measurement characteristics

Fig. 1 gives graphic illustrations [(*a*) Rasch, (*b*) 2PL] of the relationship between the measurement properties of the DSM-III-R criteria and

https://doi.org/10.1017/S0033291704003563 Published online by Cambridge University Press

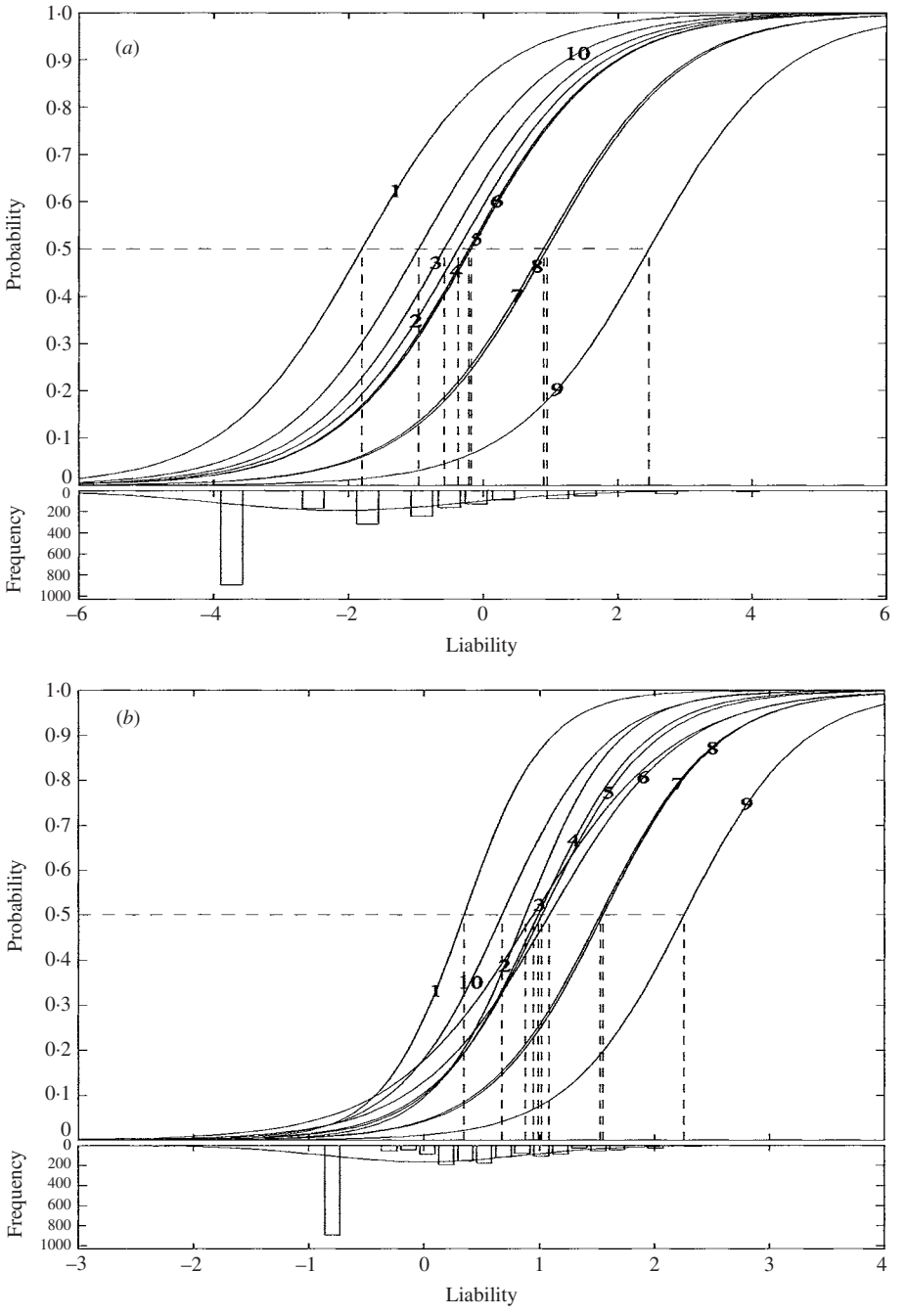

FIG. 1. DSM-III-R MD Symptom criteria calibrations and person measurement properties for the (*a*) Rasch and (*b*) 2PL models. In all latent trait models, scale origins are indeterminate and must be resolved in some manner. The Rasch model establishes a scale origin by constraining the criteria thresholds to sum to zero. Hence, 0 becomes the mean value for the criteria thresholds. Under the 2PL model, person liability levels are expressed on a linear z score-like scale (i.e. not a standardized scale) with 0 indicating 'average' risk. In practice, scores typically range between −3 and 3. Although scaling origins are arbitrary, the main feature of item-response model scaling is that both criteria and persons are jointly organized on a common unit preserving linear scale. The bottom portion of each panel presents histograms of the distributions of estimated liability to MD scores. Numbers inserted next to each IRF inflection point reference the individual MD symptom criteria as described in the text.

https://doi.org/10.1017/S0033291704003563 Published online by Cambridge University Press

corresponding person liability score distribution. Each S-shaped line is an item response function (IRF) depicting the probability of positively endorsing a symptom with increasing levels of disease liability. Two IRF properties are of primary interest. The vertical dashed lines connected to each IRF indicate the inflection point where there is a 0·5% chance of the criterion being endorsed for the corresponding level of MD liability. This location defines each criterion's liability threshold. The slope of the IRF at the inflection point indicates how strongly each criterion discriminates differences in liability. Items with steeper slopes provide sharper discriminating power as expressed by rapidly changing probabilities within small changes in liability. The steeper the IRF curve the more informative the symptom is in differentiating person risk. A key difference between the two models is that the Rasch measurement model requires all slopes to be stochastically parallel (non-overlapping) whereas the 2PL model allows slopes to vary and possibly cross over one another.

Examining the Rasch IRFs in Fig. 1(*a*), it is evident there is uneven spacing among the 10 liability thresholds with sizable gaps in measurement along the liability continuum. Compared to the other criteria, symptom 1 (depressed mood) discriminates at the lowest region of the liability continuum ($\sim -1{\cdot}8$ logits) followed by symptom 10 (>2 weeks duration) located approximately 1 logit unit above it. Likewise, symptom 9 (suicidal ideation) operates at the highest end of the liability dimension ($\sim 2{\cdot}5$ logits) with two criteria (worthlessness and difficult to concentrate) positioned within a narrow region approximately 1·5 logits below. The remaining five symptom thresholds (low interest, weight/appetite, sleeping problems, psychomotor problems, and loss of energy) are tightly grouped between $-0{\cdot}2$ and $-0{\cdot}6$ logits. These criteria all operate to differentiate risk in a narrow range suggesting possible redundancy in measurement.

In general, the 2PL threshold locations are consistent with those obtained in the Rasch model showing a similar pattern of uneven clustering. Allowing slopes to vary, some of the five clustered 'middle' criteria now have IRFs that cross over one another at various points along the liability continuum. For example, criterion 2 (low interest) and criterion 3 (weight problems) intersect in probability at $\sim 0{\cdot}75$ logits on the liability scale. Below this point, criterion 3 consistently has a higher probability of endorsement for a given risk score. However, above this cross-over point the reverse is true. This lack of a consistent relationship between the two probability curves over the scale range is an impediment to establishing a coherent interpretation of the scale as a whole. These two criteria were also flagged in the Rasch infit and outfit statistics as being statistically discrepant from model expectations.

The measurement relationships between criteria and liability scale have important implications for interpreting MD diagnoses derived from DSM symptoms. While the DSM-III-R and DSM-IV require '5 of 9' criteria to be met for a diagnosis, it is evident from these measurement results that not all combinations of five criteria define the same level of risk. Thus, there does not appear to be a straightforward correspondence between levels of risk and a positive diagnosis.

IRMs not only calibrate the DSM-III-R criteria, but also estimate person risk scores on the same scale. The lower portion of each panel in Fig. 1 displays a histogram of the distribution of estimated person liability scores with a superimposed normal curve. The very large histogram bar at the far lower end indicates that for a substantial portion of this sample, the MD diagnostic criteria provided no information to differentially assign risk scores because none of the criteria were endorsed by anyone in this group. This is an outcome often encountered when clinical survey data are collected. We note that from a measurement instrument standpoint, the DSM-III-R MD criteria are insensitive to discriminating at low levels of risk in a population-based sample.

To compare the 2PL model with one approximating a Rasch model (i.e. a 1PL model), the slopes in the 2PL model can be equated to obtain a nested model that can be formally tested. Twice the negative log likelihood is approximately distributed as chi-squared in cases where the number of items is relatively small and the model is appropriate (Bock & Aitkin, 1981). A χ^2 difference test can then be used to assess the models under normal asymptotic theory.

https://doi.org/10.1017/S0033291704003563 Published online by Cambridge University Press

Table 2. *Parameter estimates and model fits for 1- and 2-parameter logistic item response models of the 10 DSM-III-R MD criteria*

	2PL model		Equal slopes 1PL model	
Criteria name	Slope Est. (S.E.)	Threshold Est. (S.E.)	Slope Est. (S.E.)	Threshold Est. (S.E.)
1 Depressed	3·21 (0·20)	0·35 (0·03)	2·14 (0·04)	0·35 (0·04)
2 Nointerest	2·59 (0·18)	0·86 (0·04)	=	0·91 (0·04)
3 Weightapp	1·67 (0·12)	0·93 (0·05)	=	0·82 (0·04)
4 Sleepprob	2·16 (0·15)	0·99 (0·05)	=	0·98 (0·04)
5 Psycmotor	2·09 (0·15)	1·02 (0·05)	=	0·99 (0·04)
6 Fatigue	1·76 (0·13)	1·08 (0·06)	=	0·98 (0·04)
7 Worthless	1·98 (0·17)	1·54 (0·07)	=	1·47 (0·05)
8 Concentra	1·97 (0·17)	1·57 (0·08)	=	1·49 (0·05)
9 Suicidal	1·93 (0·24)	2·29 (0·15)	=	2·16 (0·07)
10 Duration	2·25 (0·15)	0·65 (0·04)	=	0·65 (0·04)
mar rel	0·71		0·71	
par est	20		11	
df	336		—	
−2 ln L	944		1014	
$\Delta\chi^2$	—		70	
Δdf	—		9	

Both the 1- and 2-PL item-response measurement models were parameterized and fit to the observed pattern frequencies in Multilog using marginal maximum likelihood. Slope Est. (S.E.) = point estimate of criteria discrimination parameter and associated standard error, Threshold Est. (S.E.) = point estimates for the liability threshold locations and corresponding standard errors.

Mar rel, marginal index of reliability; par est, number of model parameters estimated; df, degrees of freedom; –2 ln L, negative twice the log likelihood; $\Delta\chi^2$, chi-squared difference between 1PL and 2PL models; Δdf, difference in degrees of freedom.

Model-fitting results for the 2PL (columns 2–3) and a 1PL model (last 2 columns) are presented in Table 2. The equal slopes constraint produced a statistically significant increase in overall misfit as indicated by the −2 log likelihood difference for the models ($\Delta\chi^2 = 70$, Δdf = 9, $p < 0{\cdot}001$). This test indicates the 1PL model does a poorer job of accounting for the observed marginal frequencies compared to the less restrictive 2PL model. However, relying solely on a single valued misfit statistic to reject models should be done with care. Model misfit statistics are sensitive to and can be affected by: (1) sample size, (2) sparseness of observed response patterns, (3) the stringent statistical criteria imposed by exact tests of metric invariance, (4) person misfit, and (5) population heterogeneity (Muthén, 1989*b*). Therefore, given the reasonable Rasch results and the implications this model has for interpreting symptom counting, we proceed to consider both IRM scales when examining external predictive validity.

Validity of IRM scales *versus* DSM-III-R diagnosis

Table 3 presents linear and logistic regression results predicting neuroticism (N) and a later diagnosis of MD for a sequence of models with, (1) an intercept only, (2) adding Age, (3) the binary diagnostic variable (Dx), (4) the quantitative liability variable (Risk), and (5) a Dx × Risk interaction (Dx-Risk). The interaction term is included for comparison completeness.

Results for the N score outcome are given in the upper portion and those for following year MD diagnosis in the lower portion of Table 3. The coefficient changes found between models 3 and 4 are of primary interest. With only Age and Dx in the model, the raw Dx coefficient value of 2 is statistically significant ($\chi^2 = 75$, df = 1, $p = 0{\cdot}0001$). This is the estimated mean N difference score between non-cases (scored 0) and cases (scored 1). Coefficients are noticeably altered when the quantitative liability scale is added in model 4. The quantitative risk coefficient is 0·51 ($\chi^2 = 118$, df = 1, $p = 0{\cdot}0001$) whereas the effect of Dx vanishes (0·03; $\chi^2 = 0{\cdot}01$, df = 1, $p = 0{\cdot}93$). A comparison of the standardized coefficients indicates a substantial difference in effect size between the quantitative risk (0·29) and binary Dx (0·002) variables although such comparisons need to be interpreted with caution (Greenland *et al*. 1986).

An examination of the partitioned sums of squares further demonstrates the superior statistical power of the IRM risk variable. Contrasting the Type I (sequential) sums of squares (Age = 87, Dx = 767, Risk = 1144, Dx-Risk = 0·1) with the Type II (simultaneous) sums of squares (Age = 59, Dx = 0·07, Risk = 1144, Dx-Risk = 0·1), it is evident that the added variability in the quantitative risk scale overpowers the binary mean difference distinction. Similar results were found for the 2PL quantitative risk scale with the noted exception that in the full model (5), the sign of the Dx coefficient is reversed (−0·19). The raw regression coefficients are expressed in different units due to the different scale anchoring used in the Rasch and 2PL estimation procedures.

The lower portion of Table 3 presents logistic regression results for predicting a diagnosis of MD in the following year. The quantitative risk scales again dominate when added to the

Table 3. *Parameter estimates and associated tests of statistical significance for a sequential series of regression models predicting neuroticism total scores* (*upper portion*) *and following year MD diagnosis* (*lower portion*)

	Parameter estimate				
Int	Age	Dx	Risk	Dx-Risk	
Neuroticism					
Rasch model					
1	5·7 (0) [6442, 0·0001]	—	—	—	—
2	6·6 (0) [502, 0·0001]	−0·03 (−0·06) [8·2, 0·004]	—	—	—
3	6·4 (0) [497, 0·0001]	−0·03 (−0·07) [10, 0·002]	**2·0 (0·19) [75, 0·0001]**	—	—
4	7·5 (0) [637, 0·0001]	−0·02 (−0·05) [6·1, 0·01]	**0·03 (0·002) [0·01, 0·93]**	**0·51 (0·29) [118, 0·0001]**	—
5	7·5 (0) [632, 0·0001]	−0·02 (−0·05) [6·0, 0·01]	0·06 (0·005) [0·02, 0·88]	0·52 (0·29) [113, 0·0001]	−0·03 (−0·004) [0·01, 0·91]
2PL model					
4	6·3 (0) [500, 0·0001]	−0·02 (−0·05) [6·0, 0·01]	0·04 (0·004) [0·02, 0·89]	1·3 (0·29) [119, 0·0001]	—
5	6·3 (0) [500, 0·0001]	−0·02 (−0·05) [6·0, 0·01]	−0·19 (0·004) [0·03, 0·85]	1·3 (0·29) [115, 0·0001]	0·16 (0·02) [0·05, 0·82]
Following year MD diagnosis					
Rasch model					
1	−2·2 (—) [870, 0·0001]	—	—	—	—
2	−2·8 (—) [74, 0·0001]	0·01 (0·05) [1·1, 0·21]	—	—	—
3	−2·5 (—) [74, 0·0001]	0·01 (0·04) [1·0, 0·30]	**1·5 (0·24) [64, 0·0001]**	—	—
4	−1·8 (—) [29, 0·0001]	0·02 (0·07) [2·5, 0·12]	**−0·50 (−0·08) [3·6, 0·06]**	**0·57 (0·58) [91, 0·0001]**	—
5	−1·8 (—) [88, 0·0001]	0·02 (0·07) [2·4, 0·12]	−0·31 (−0·05) [0·95, 0·89]	0·59 (0·60) [83, 0·0001]	−0·17 (−0·05) [0·96, 0·33]
2PL model					
4	−3·2 (—) [88, 0·0001]	0·02 (0·07) [2·4, 0·12]	−0·49 (−0·08) [0·02, 0·89]	1·4 (0·59) [90, 0·0001]	—
5	−3·2 (—) [88, 0·0001]	0·02 (0·07) [2·4, 0·12]	−0·41 (−0·07) [0·26, 0·61]	1·4 (0·59) [82, 0·0001]	−0·06 (−0·01) [0·01, 0·91]

Int, intercept only model; Age, adding age in years; Dx, adding binary MD diagnosis variable; Risk, adding continuous item-response theory measured risk; Dx-Risk, adding diagnosis × risk interaction.

Raw regression coefficients, standardized values ( ), and χ^2 values with corresponding exact probability for a 1 degree of freedom statistical significance test [ ] are displayed for each parameter in the model.

regression model (models 4 and 5). Although the Dx variable is a significant predictor by itself (model 3: 1·5, $\chi^2=64$, df = 1, $p=0·0001$), it changes signs and is rendered non-significant when the risk scale is added (model 4: −0·5, $\chi^2=3·6$, df = 1, $p=0·06$). The risk scale coefficient estimate is 0·57 ($\chi^2=91$, df = 1, $p=0·0001$) and standardized values again suggest a large effect size difference (−0·08 for Dx compared to 0·58 for Risk).

Twin correlations and heritability estimates obtained from fitting a standard twin model are shown in Table 4. Parameters are estimated for the two IRM risk scales under continuous and ordinal (labeled accordingly in Table 4) scaling. This is to check for possible failure of the multivariate normality assumption. Treating Rasch scores as ordinal ignores their equal interval properties. For the 2PL scale, the 321 unique risk values were collapsed into 15 ranked categories for the ordinal analysis.

Twin correlation point estimates were slightly higher for the MD diagnosis variable but did not significantly differ from the Rasch and 2PL liability scale estimates. However, statistical precision for the IRM scale correlation estimates was noticeably improved. For example, the span of the 95% CI region for the binary MZ correlation was 0·36 but only 0·13/0·16 for the LT-IRM liability scales modeled as continuous and ordinal variables respectively.

In the twin models, genetic and environmental parameter CIs were narrower for the risk scales. The additive genetic proportion of variance estimate for the DSM diagnosis variable, even with a point estimate of 0·42, could not be statistically distinguished from zero as indicated by the lower CI of 0·000. For the IRM risk variables, point estimates of additive genetic proportions of variance were slightly lower but the lower CIs did not include zero (0·19/0·13 and 0·18/0·17 for the Rasch and 2PL models under continuous and ordinal estimation respectively). The total 95% CI range was 0·58 for the binary diagnosis compared to only 0·24 for the LT-IRM scale. Thus, the CI regions for the IRM scales were ∼35–40% smaller than those obtained for the binary MD diagnosis

https://doi.org/10.1017/S0033291704003563 Published online by Cambridge University Press

Table 4. *Comparison of twin correlations and heritability components of variance for IRM-derived risk scales and a binary DSM diagnostic variable*

	Rasch model			2PL model			DSM-III-R diagnosis		
	(LCI)	Est.	(UCI)	(LCI)	Est.	(UCI)	(LCI)	Est.	(UCI)
					Twin correlations				
Continuous									
MZ correlation	(0·290)	0·359	(0·423)	(0·295)	0·364	(0·428)	—	—	—
DZ correlation	(0·145)	0·179	(0·247)	(0·148)	0·182	(0·253)	—	—	—
Ordinal									
MZ correlation	(0·301)	0·383	(0·457)	(0·299)	0·379	(0·452)	(0·221)	0·418	(0·582)
DZ correlation	(0·153)	0·192	(0·289)	(0·150)	0·189	(0·272)	(0·114)	0·208	(0·402)
					Heritability (components of variance)				
Continuous									
Additive genetic	(0·186)	0·359	(0·423)	(0·183)	0·364	(0·428)	—	—	—
Common environment	(0·000)	0·000	(0·141)	(0·000)	0·000	(0·148)	—	—	—
Unique environment	(0·577)	0·641	(0·710)	(0·572)	0·636	(0·705)	—	—	—
Ordinal									
Additive genetic	(0·132)	0·382	(0·457)	(0·167)	0·379	(0·453)	(0·000)	0·416	(0·582)
Common environment	(0·000)	0·000	(0·210)	(0·000)	0·000	(0·175)	(0·000)	0·000	(0·399)
Unique environment	(0·543)	0·618	(0·699)	(0·547)	0·621	(0·700)	(0·418)	0·584	(0·779)

MZ and DZ correlations, parameter estimates and confidence intervals were obtained using full information raw maximum likelihood as implemented in the Mx software. (LCI), estimated lower 95% confidence interval; (UCI), estimated upper 95% confidence interval. Additive genetic, common environment, and unique environment parameter estimates are expressed as proportions of variances.

variable. Given that the CI decreases as a function of the square of the sample size, to obtain the same parameter accuracy with the binary MD diagnosis variable, a sample *6 times larger* would be required Neale *et al*. 1994.

DISCUSSION

In applying IRMs to MD symptom data collected in a community sample, we sought first to investigate how the DSM-III-R criteria for MD performed under a quantitative measurement approach. A second interest was to determine what could be gained in the way of additional information if symptom response data were utilized to construct a continuous scale of liability rather than create a dichotomous diagnostic variable. We examine these two research objectives in turn.

Measuring properties of DSM-III-R MD criteria

Do the DSM-III-R MD criteria define a coherent unitary dimension of liability? Exploratory and confirmatory factor analysis and Rasch model results suggest that, as a first approximation, the 10 criteria worked reasonably well together to define a single dimension of liability to MD. Given that the DSM criteria sets for MD, whose history can be traced back to the Feighner and Research Diagnostic Criteria (Feighner *et al*. 1972; Spitzer *et al*. 1975), were originally developed by 'expert' clinicians with no intentions to measure risk, the degree to which the MD criteria conformed to the requirements of fundamental measurement were unexpected. However, when the disaggregated set of 14 DSM depressive symptom criteria were examined, the unidimensionality condition was much less tenable with results being more consistent with prior population-based evidence suggesting a typical *versus* atypical depressive distinction (Horwath *et al*. 1992; Kendler *et al*. 1996).

Related questions examined how well the 10 DSM criteria defined a common scale of MD liability and if the criteria set showed even measurement across the risk continuum. The results across the analytical methods used were consistent. The weight/appetite and fatigue criteria performed less well than the others. These criteria showed poorer discrimination, greater misfit and, in general, were less useful than the other criteria in defining a coherent dimension of liability. The loss of interest, weight problems, sleep problems, psychomotor problems, and fatigue criteria all tended to discriminate risk within a narrow scale range. Three of these

https://doi.org/10.1017/S0033291704003563 Published online by Cambridge University Press

criteria (i.e. weight, sleep, and psychomotor problems) were collapsed into binary polar variants prior to the IRM analyses.

A final research interest was to investigate the properties of the person liability scores constructed under the IRMs. Good construct measurement validity depends on how well a particular set of criteria differentiates person risk across the range of liability. In this respect, the performance of the DSM criteria was less clear. The majority of criteria tended to operate at the upper half of the liability scale and, in terms of measuring risk, were 'off-target' with respect to liability in this population-based sample of female twins. Criteria were insensitive to distinguishing differences at low levels of risk. The marginal reliability – an index of the average reliability across all risk levels – was modest (0·71, see Table 2).

Validity of the IRM liability estimates

A major advantage of IRMs is that a liability score is estimated for each individual in the sample. If MD liability is well represented by a continuous distribution, variability will be grossly restricted and possibly distorted when individuals are classified into mutually exclusive cases *versus* non-cases. Using three external validity criteria, the IRM liability scales and standard MD diagnosis variables were jointly examined to determine their respective predictive power. IRM-derived liability scales clearly out-performed the binary diagnosis variable when both were included in linear and logistic prediction models. When used in genetic models, the IRM liability scales substantially improved the estimation precision of the heritability parameters. Additive genetic effects were found to be significantly different from zero for the liability scales but not for the binary diagnosis. This was despite the fact that the point estimate was *larger* for the binary diagnosis than for the continuous liability scales.

The findings reported here suggest that for many research purposes, available information in symptom data is not being fully utilized. In certain circumstances (e.g. estimation of prevalence or selection into controlled treatment trials), a dichotomous MD diagnosis may be necessary. But for research into the prediction of outcomes, the delineation of important risk factors, and a more detailed examination into relationships between psychiatric phenotype variants and genetic architecture, the aggregation of DSM criteria data into dichotomies may restrict and limit such efforts.

The LT-IRM liability scaling results are directly relevant to the practice of counting symptoms. If the Rasch measurement model requirements are satisfied, the summed score is a sufficient statistic for independently estimating criteria and person parameters. This evidence lends some support to the DSM diagnostic practice of counting symptoms and the use of such a count as a crude proxy for risk level. These results also indicate whether a symptom count is an appropriate outcome variable for statistical analysis. However, as important as the properties of the Rasch model are to establishing sound measurement, the premise that all non-essential symptom criteria contribute equally to determining a diagnosis may not be plausible from a clinical perspective. For example, an individual who endorsed only the depressed mood symptom (lowest risk threshold) and another who only endorsed the suicidal ideation symptom (highest risk threshold) would have the same marginal symptom count (i.e. 1) and hence be given the same risk score. This would likely conflict with the interpretation attributed to the clinical impact of each symptom.

The difference between the latent liability scales constructed under the LT-IRM approach and the theoretical latent variable typically introduced when analyzing a binary diagnostic variable is also important. In the analysis of the diagnostic variable, a single threshold is estimated on the hypothetical latent liability variable from the proportion of cases *versus* non-cases. This inferred latent variable is usually assumed to be continuous and normally distributed. All the available symptom level information is collapsed into this single unaffected/affected threshold.

Using the LT-IRM approach, all the criterion level information is drawn upon to simultaneously estimate values and standard errors for criteria and person parameters. The LT-IRM scaling procedure constructs a continuous index of phenotypic risk using all of the symptom endorsement patterns. In contrast, the latent variable in the binary analysis only makes contact with observed data via the single affect

https://doi.org/10.1017/S0033291704003563 Published online by Cambridge University Press

versus unaffected threshold. Hence, in many research contexts the LT-IRM approach may serve as a more appropriate way to represent self-report data than are the clinician-like binary classification decisions that result from applying diagnostic algorithms.

Finally, the analyses presented here only begin to exploit the potential empirical and conceptual richness of IRMs when applied to psychiatric symptom survey data. For example, it is possible to develop criteria 'profiles' for each individual, thereby identifying various types of person misfit and subjects with unusual endorsement patterns. Such 'outlier' profiles may reflect rare subforms of illness, misunderstood questions or subject misrepresentation. Measurement approaches also make it possible to take into account relevant auxiliary information collected on symptoms. For example, rather than simply coding symptoms as either present or absent, symptom impact can be encoded in a graded fashion (e.g. 0 = not present, 1 = present, mild, 2 = present, severe).

Limitations

These results should be interpreted in the context of several considerations. Although IRT, and especially the Rasch model, provide a theoretical framework for estimating item and person measurement properties independent of the items and samples used to obtain them, this is a community sample and results may differ in treated samples. We have studied only white female twins born in Virginia. These findings may or may not extrapolate to other populations or to men.

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 the ascertainment of subjects for this study. The MATR, directed by Dr J. Silberg and Dr L. Eaves, has received support from the National Institutes of Health, the Carman Trust and the WM Keck, John Templeton and Robert Wood Johnson Foundations. Indrani Ray provided database assistance.

DECLARATION OF INTEREST

None.

REFERENCES

APA (1987). *Diagnostic and Statistical Manual of Mental Disorders* (3rd edn, revised). American Psychiatric Association. Washington, DC.

Andrich, D. (1989). Distinctions between assumptions and requirements in measurement in the social sciences. In *Mathematical and Theoretical Systems* (ed. J. A. Keats, R. Taft, R. A. Heath and S. H. Lovibond), pp. 7–16. North-Holland: New York.

Angst, J. & Clayton, P. (1986). Premorbid personality of depressive, bipolar, and schizophrenic patients with special reference to suicidal issues. *Comprehensive Psychiatry* **27**, 511–532.

Bartholomew, D. J. (1980). Factor analysis for categorical data. *Journal of the Royal Statistical Society, Series B* **42**, 293–321.

Bell, R. C., Low, L. H., Jackson, H. J., Dudgeon, P. L., Copolov, D. L. & Singh, B. S. (1994). Latent trait modelling of symptoms of schizophrenia. *Psychological Medicine* **24**, 335–345.

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

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In *Statistical Theories of Mental Test Scores* (ed. F. M. Lord and M. R. Novick), pp. 397–422. Addison-Wesley: Reading, MA.

Blashfield, R. K. & Livesley, W. J. (1991). Metaphorical analysis of psychiatric classification as a psychological test. *Journal of Abnormal Psychology* **100**, 262–270.

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

Browne, M. W. & Cudeck, R. (1993). Alternative ways of assessing model fit. In *Testing Structural Equation Models* (ed. K. S. Bollen and J. S. Long), pp. 136–162. Sage: Newbury Park, CA.

Christoffersson, A. (1975). Factor analysis of dichotomous variables. *Psychometrika* **40**, 5–32.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement* **7**, 249–253.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* **39**, 1–38.

Duncan-Jones, P., Grayson, D. A. & Moran, P. A. P. (1986). The utility of latent trait models in psychiatric epidemiology. *Psychological Medicine* **16**, 391–405.

Embretson, S. E. & Reise, S. P. (2000). *Item Response Theory for Psychologists*. Lawrence Erlbaum Associates: New York.

Faraone, S. V. & Tsuang, M. T. (1994). Measuring diagnostic accuracy in the absence of a 'gold standard'. *American Journal of Psychiatry* **151**, 650–657.

Feighner, J. P., Robins, E., Guze, S. B., Woodruff, R. A., Winokur, G. & Munoz, R. (1972). Diagnostic criteria for use in psychiatric research. *Archives of General Psychiatry* **26**, 57–63.

Gibbons, R. D., Clark, D. C. & Kupfer, D. J. (1993). Exactly what does the Hamilton Depression Rating Scale measure? *Journal of Psychiatric Research* **27**, 259–273.

Gibbons, R. D., Clark, D. C., VonAmmon, C. S. & Davis, J. M. (1985). Application of modern psychometric theory in psychiatric research. *Journal of Psychiatric Research* **19**, 43–55.

Goldstein, H. & Wood, R. (1989). Five decades of item response modelling. *British Journal of Mathematical and Statistical Psychology* **42**, 139–167.

Grayson, D. A., Bridges, K., Duncan-Jones, P. & Goldberg, D. P. (1987). The relationship between symptoms and diagnoses of minor psychiatric disorder in general practice. *Psychological Medicine* **17**, 933–942.

Greenland, S., Schlesselman, J. J. & Criqui, M. H. (1986). The fallacy of employing standardized regression coefficients and correlations as measures of effect. *Journal of Epidemiology* **123**, 203–208.

Hirschfeld, R. M. A., Klerman, G. L., Lavori, P. W., Keller, M. B., Griffith, P. & Coryell, W. (1989). Premorbid personality assessments of first onset of major depression. *Archives of General Psychiatry* **46**, 345–350.

Horwath, E., Johnson, J., Weissman, M. M. & Hornig, C. D. (1992). The validity of major depression with atypical features based on a community study. *Journal of Affective Disorders* **26**, 117–126.

Kendler, K. S. (1990). Toward a scientific psychiatric nosology: strengths and limitations. *Archives of General Psychiatry* **47**, 969–973.

Kendler, K. S., Eaves, L. J., Walters, E. E., Neale, M. C., Heath, A. C. & Kessler, R. C. (1996). The identification and validation of distinct depressive syndromes in a population-based sample of female twins. *Archives of General Psychiatry* **53**, 391–399.

Kendler, K. S., Neale, M. C., Kessler, R. C., Heath, A. C. & Eaves, L. J. (1993). A longitudinal twin study of personality and major depression in women. *Archives of General Psychiatry* **50**, 853–862.

Lewine, R. J., Fogg, L. & Meltzer, H. Y. (1983). Assessment of negative and positive symptoms in schizophrenia. *Schizophrenia Bulletin* **9**, 368–376.

Linacre, J. M. & Wright, B. D. (2000). *A User's Guide to Winsteps Rasch-Model Computer Programs*. MEAS Press: Chicago, IL.

Lord, F. M. & Novick, M. R. (1969). *Statistical Theories of Mental Test Scores*. Addison Wesley: Reading, MA.

MacCallum, R. C., Zhang, S., Preacher, K. J. & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods* **7**, 19–40.

Marsh, H. W., Balla, J. R. & McDonald, R. P. (1988). Goodness of fit indexes in confirmatory factor analysis: the effect of sample size. *Psychological Bulletin* **103**, 391–410.

Mitchell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology* **88**, 355–383.

Muthén, B. O. (1989*a*). Dichotomous factor analysis of symptom data. *Sociological Methods & Research* **18**, 19–65.

Muthén, B. O. (1989*b*). Latent variable modeling in heterogeneous populations. *Psychometrika* **54**, 557–585.

Muthén, B. O. & Muthén, L. K. (eds.) (2001). *Mplus User's Guide*. Muthen & Muthen: Los Angeles, CA.

Neale, M. C., Boker, S. M., Xie, G. & Maes, H. H. (1999). *Mx: Statistical Modeling* (5th edn). Department of Psychiatry, Medical College of Virginia, Commonwealth University, Box 980126, Richmond, VA 23298.

Neale, M. C., Eaves, L. J. & Kendler, K. S. (1994). The power of the classical twin study to resolve variation in threshold traits. *Behavior Genetics* **24**, 239–258.

Orlando, M., Sherbourne, C. D. & Thissen, D. (2000). Summed-score linking using item response theory: application to depression measurement. *Psychological Assessment* **12**, 354–359.

Pickles, A. & Angold, A. (2003). Natural categories or fundamental dimensions: On carving nature at the joints and the rearticulation of psychopathology. *Developmental Psychopathology* **15**, 529–551.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment teats. Danish Institute for Educational Research, Copenhagen.

Rasch, G. (1966). An item analysis which takes individual differences into account. *Journal of Mathematical and Statistical Psychology* **19**, 49–57.

Reiser, M. (1989). An application of the item-response model to psychiatric epidemiology. *Sociological Methods & Research* **18**, 66–103.

Robins, E. & Guze, S. B. (1970). Establishment of diagnostic validity in psychiatric illness: Its application to schizophrenia. *American Journal of Psychiatry* **126**, 983–987.

Spitzer, R. L., Endicott, J. & Robins, E. (1975). *Research Diagnostic Criteria for a Selected Group of Functional Disorders* (2nd edn). New York Psychiatric: New York.

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

Steiger, J. H. (1990). Structural model evaluation and modification: an interval estimation approach. *Multivariate Behavioral Research* **25**, 173–180.

Thissen, D. (1991). *Multilog Users' Guide: Multiple Categorical Item Analysis and Test Scoring using Item Response Theory*. Scientific Software, Inc.: Chicago, IL.

Wainwright, N. W. J., Surtees, P. G. & Gilks, W. R. (1997). Diagnostic boundaries, reasoning and depressive disorder, I: Development of a probabilistic morbidity model for public health psychiatry. *Psychological Medicine* **27**, 835–845.

Wright, B. D. & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions* **8**, 370

Wright, B. D. & Stone, M. H. (1979). *Best Test Design*. MESA Press: Chicago, IL.