Journal of Abnormal Child Psychology, Vol. 34, No. 3, June 2006, pp. 379–391 (© 2006)
DOI: 10.1007/s10802-006-9027-x

The Short Mood and Feelings Questionnaire (SMFQ): A Unidimensional Item Response Theory and Categorical Data Factor Analysis of Self-Report Ratings from a Community Sample of 7-through 11-Year-Old Children

Carla Sharp,[1,4] Ian M. Goodyer,[2] and Tim J. Croudace[3]

Received February 24, 2004; revision received May 10, 2005; accepted September 13, 2005
Published online: 29 April 2006

Item response theory (IRT) and categorical data factor analysis (CDFA) are complementary methods for the analysis of the psychometric properties of psychiatric measures that purport to measure latent constructs. These methods have been applied to relatively few child and adolescent measures. We provide the first combined IRT and CDFA analysis of a clinical measure (the Short Mood and Feelings Questionnaire—SMFQ) in a community sample of 7-through 11-year-old children. Both latent variable models supported the internal construct validity of a single underlying continuum of severity of depressive symptoms. SMFQ items discriminated well at the more severe end of the depressive latent trait. Item performance was not affected by age, although age correlated significantly with latent SMFQ scores suggesting that symptom severity increased within the age period of 7–11. These results extend existing psychometric studies of the SMFQ and confirm its scaling properties as a potential dimensional measure of symptom severity of childhood depression in community samples.

KEY WORDS: Screening; childhood depression; SMFQ; item response theory; categorical data factor analysis.

Over the last 40 years, the methods used to evaluate the psychometric basis of ability tests, health care surveys, and multi-item screening instruments has changed dramatically. Whilst the methodology of classical test theory (CTT) has served test development well, item response/latent trait theory (IRT) approaches have become more mainstream as the technical basis for measurement theory, test construction and scale evaluation (Embretson & Reise, 2000). Although moves towards adoption of more appropriate, non-linear and categorical data factor analysis (CDFA) models have been most apparent in educational settings, in the last two decades such methods have begun to be applied in clinical testing of adults. This has been evidenced by psychometric studies published in, for example, *Psychological Assessment* and *Psychological Methods* (e.g. Angold, Erkanli, Silberg, Eaves, & Costello, 2002; Cooke & Michie, 1997; Lambert et al., 2003; Patton, Carlin, Shao, Hibbert, & Bowes, 1997; Santor, Ramsay, & Zuroff, 1994). Currently there are very few reports that have applied such methodologies in samples of young children (Cheong & Raudenbush, 2000).

One reason for the under-exploitation of such methodologies may be because researchers have not been introduced to the potential and practicalities of these methods (Rouse, Finger, & Butcher, 1999) and are therefore unaware of the advantages they offer over conventional (CTT) methods (van der Linden & Hambleton, 1997). Although CTT is often included in the curriculum of both clinical and applied psychologists, IRT is rarely taught, and has had less coverage in mainstream

[1] Menninger Department of Psychiatry and Behavioral Sciences, Baylor College of Medicine, Houston, Texas.
[2] Developmental Psychiatry Section, University of Cambridge, Cambridge, UK.
[3] Department of Psychiatry, University of Cambridge, Cambridge, UK.
[4] Address all correspondence to Carla Sharp, Menninger Department of Psychiatry and Behavioral Sciences, Baylor College of Medicine, 1 Baylor Plaza, Houston, Texas 77030; e-mail: csharp@hnl.bcm.tmc.edu.

psychology journals (Embretson & Reise, 2000). We provide an application of IRT and categorical data factor analysis (CDFA) methods to a commonly used self-report measure of depressive symptoms in children, the Short Mood and Feelings Questionnaire (SMFQ; Angold et al., 1995). As such, the aim of this paper was to scrutinize the internal construct validity of the SMFQ. To this end, we used latent variable models implemented with both an IRT and appropriate factor analysis framework (CDFA). Within a latent variable framework, internal construct validity refers to an understanding of the psychometric performance of items in relation to an underlying (latent) construct of interest. Here the latent variable was a continuum of severity derived from self-reported depressive symptoms in 7–11-year-old children. This latent variable was psychometrically derived, and was not validated through an external criterion measure of depression; hence our results pertain only to internal construct validity.

The Mood and Feelings Questionnaire (MFQ), long form, was developed as a screening tool for detecting clinically meaningful signs and symptoms of depressive disorders in children and adolescents (6–17 years of age) by self-report (Angold et al., 1995; Costello & Angold, 1988). MFQ items were designed to cover DSM diagnostic criteria for major depressive disorder (APA, 1994). Over the past decade, the long form consisting of 33 items has been used extensively in both epidemiological and clinical research (Costello et al., 1996a, 1996b; Goodyer, Herbert, Tamplin, & Altham, 2000; Kent, Vostanis, & Feehan, 1997; Park, Goodyer, & Teasdale, 2002; Wood, Kroll, Moore, & Harrington, 1995). Criterion-related validity (ability to predict clinical diagnosis) has been established for the long form (Wood et al., 1995).

A short form consisting of 13 items (SMFQ) was subsequently derived for which criterion validity has also been shown (Angold et al., 1995; Kent et al., 1997; Thapar & McGuffin, 1998). Table I presents details of studies that have used the self-report version of both the long and short form of the MFQ and summarizes the validity issues addressed. Although these studies are universally supportive of the internal construct validity of the SMFQ, most samples were from clinic or specially selected populations. Currently, there is no published study on the internal construct validity of the short version in a community sample of primary school-aged children. Establishing internal construct validity in 7–11-year-old children is essential if the SMFQ is to be recommended as a self-report screening measure of severity of depression in community samples of children.

Most existing studies, as summarized in Table I, have applied statistical procedures based on CTT, linear factor analysis or principal components. Such methods are not optimal for the discrete/categorical nature of the MFQ responses (Angold et al., 2002) for several reasons. Traditional methods, such as principal components analysis, assume that item responses are on a continuous metric, yet, psychopathology ratings are recorded using discrete categories in most community studies to date. Usually, the commonest (modal) response on SMFQ items is zero, representing absence of symptoms. Even when ratings are collected on graded scales, response distributions are usually heavily skewed. Failure to treat symptom ratings as categorical data in factor analysis models has two consequences: (1) in multi-dimensional analysis the true factor structure may be severely distorted, and (2) in unidimensional models factor loadings may be biased and resultant (weighted) scores estimated incorrectly. When applying linear models to binary (0,1) data, predictions are made that are not within the plausible range (<0 or >1) (McDonald, 1999). This is obviously highly undesirable when relatively rare symptoms are being modelled. The limitations of linear models in these situations are well known to statisticians and methodologists, but are frequently ignored by applied researchers. Only in limited circumstances will linear methods approximate a more appropriate model (Shrout & Parides, 1992).

Although CTT and IRT/CDFA both assume that variation in the observed responses to items of a test can be explained by one or more continuous unobserved latent traits, the way IRT/CDFA models the relation between observed item-responses and the latent trait is different. Instead of summarizing the psychometric properties of a scale with omnibus statistics (such as item-total correlations or Cronbach alpha), thereby averaging across levels of individual variation (Santor et al., 1994), IRT approaches model how the probability of responding to an item—here this is equivalent to endorsing a symptom—varies as a function of the location along a latent continuum or dimension of variation (Santor et al., 1994). IRT methods do not use summary statistics that apply to groups of individuals, such as correlations, but can define a model for the individual response patterns that comprise the raw data. Because item response patterns can be modelled directly within an IRT framework, no information in the data is lost.

A mathematical equation—a probability model, similar to that used in logistic or probit regression—is used to describe the non-linear relation between an item-response and the value (severity) on the latent trait. This relation can be represented graphically by a plot that is known as an item characteristic curve (ICC) or item response function (IRF). The ICC offers a graphical profile of item

Table I. Summary of Validity Characteristics of Studies that Used the Self-report Versions of the MFQ Long and Short Forms

Author and date	Samples	Age	*N* Sex	Type of validity assessed	Analytic strategy	Results
Wood et al. (1995)	104 consecutive referrals to outpatient psychiatric clinic	10–19	43 boys 61 girls	Criterion validity for long form, using the K-SADS as criterion	ROC analyses	Sensitivity 0.78 Specificity 0.78
Angold et al. (1995)	48 consecutive referrals to outpatient psychiatric clinic	6–17	33 boys 15 girls	Internal consistency, content and criterion-related validity for SMFQ, using the DISC as criterion	Principle component analysis Cronbach's alpha	Unidimensional factor structure (acceptable model fit statistics) Cronbach's alpha 0.85
	125 referrals to a pediatric clinic	6–11	54 boys 71 girls		Maximum likelihood logistic regression ROC analyses	High ORs for predicting psychiatric diagnosis Sensitivity 0.60 Specificity 0.85
Messer et al. (1995)	1502 high risk community children	6–13	Boys only	Internal consistency	CFA using LISREL VII	Unidimensional structure of SMFQ confirmed: GFI and AGFI indices high (>0.97); RMSR low (>0.08); χ^2/df indices all <3
Kent et al. (1997)	114 consecutive attendees at four clinics	7–17	56 boys 57 girls	Criterion validity for long and SMFQ, using the K-SADS as criterion	Correlational analyses ROC analyses	Sensitivity 0.59 Specificity 0.89
Thapar and McGuffin (1998)	411 twins	11–16	99 boy pairs 123 girl pairs 94 mixed pairs	Criterion validity for SMFQ, using the CAPA as criterion	Correlational analyses ROC analyses	Sensitivity 0.75 Specificity 0.74

Note. Criterion validity here refers to the SMFQ's ability to detect major depressive episode in study samples with an acceptable degree of sensitivity and specificity. Internal consistency refers to the underlying factor structure of the SMFQ, notably whether it can be demonstrated to be a unidimensional scale. The current study did not aim to examine criterion validity but only internal consistency.

effectiveness, which is more informative than traditional measures of item performance (Santor et al., 1994). For a detailed discussion of the distinction between CTT and IRT approaches and the benefits of using the latter, see Embretson and Reise (2000).

More detailed information on the effective measurement range of individual items and scales is especially important for psychopathology measurement in developmental epidemiology studies. Epidemiological evidence has suggested that symptom profiles may differ with age for reasons that are not fully understood. Weiss and Garber (2003) outlined several ways in which developmental differences may impact on the phenomenology of depression over the course of childhood and adolescence. From a psychometric perspective, it may be, for instance, that certain items are not appropriate for certain age groups, or that the symptoms (defined by the wording of questionnaire items) are not a feature of a particular disorder in that age group. As the current study demonstrated (see section on differential item function—DIF), IRT provides the opportunity to distinguish between bias at the level of the item (i.e. the item does not accurately probe for the symptom for a particular age group) and bias at the level of the latent trait (i.e. the disorder does not express itself through a particular symptom in a particular age group).

To our knowledge, only one study to date has adopted an appropriate categorical factor analysis

approach to SMFQ data (Angold et al., 2002). Angold et al. (2002) reported CDFA solutions estimated using weighted least squares estimation methods in Mplus software. Their study confirmed the SMFQ's unidimensional structure (single factor) in two samples: First, the Great Smoky Mountains Study comprised 9-, 11- and 13-year-olds. This sample of $n = 1441$ represented 25% of the highest scorers on the Child Behaviour Checklist out of a community-based sample of $n = 4500$, i.e. a "high-risk" sample. Similar results were found in a second sample from a family study of $n = 1412$ twins.

Confirming the unidimensional factor structure of the SMFQ with non-linear factor analytic techniques (Angold et al., 2002) was an important first step in examining the psychometric properties of the SMFQ from a latent trait modelling perspective. However, no information was presented on the effective measurement range of the SMFQ. This is more easily summarised using graphical representations of the model from an IRT perspective, i.e. the test characteristic curve (TCC). We repeated and extended the methods used by Angold et al. and offer more graphic representations of the latent variable modelling. In addition, we provide the first IRT/CDFA modelling of SMFQ data in a community sample of younger children (including 7- and 8-year-olds).

Given concerns that there might be item bias with respect to age (with different thresholds for response for younger and older children), we further extended Angold's approach by testing for item bias (DIF) using a multiple indicator multiple cause (MIMIC) modelling approach (Gallo, Anthony, & Muthén, 1994). Item bias is present when individuals with the same score on the psychometrically derived latent trait are more or less likely to endorse an item. MIMIC modelling investigates item bias through an extension of the CDFA factor analysis model to include covariates, both of the latent trait, and of the items. In testing for age invariance of item thresholds, the estimates of interest are the direct effects (regressions of) item responses on age as a covariate after adjustment for the effect of age on the latent trait score. Given that prevalence rates of depression increase with age in adolescence (Goodyer & Sharp, 2005; Hankin et al., 1998), we expected this might also be the case for this age group. We therefore anticipated finding an age difference in latent trait scores, indicating more psychopathology in the older children (10- and 11-year-olds), but no evidence of item bias for any of the SMFQ items. We were able to test for the effect of age, because unlike previously reported studies, our sample comprised a cohort of younger children (7–11-year-olds) recruited and assessed in an elementary school setting.

METHOD

Participants

Parents of 2950, 7- to 11-year-old children (primary school years 3–6) of 16 primary schools from a mixed catchment of rural and urban areas in Cambridgeshire, England were asked to participate. Response rates for individual schools ranged from 14 to 40% resulting in 20% of the children taking part in the study ($n = 659$; 319 boys and 340 girls).

There are four possible reasons to explain the low response rate. First, the ethics approval requirements prohibit researchers from gaining access to names and addresses of parents in the community. Invitation letters to participate in the study were therefore handed out to children at school to take home to their parents. It is possible that many invitation letters did not make it home in the first place. Second, ethics in the UK require positive consent. The effort of actually completing and returning a consent form to indicate positive consent may be too much to ask of some parents in the community. Third, limited resources precluded payment to children for their participation. Instead, children were given a sticker and were entered into a school raffle drawing for their participation. Fourth, it is possible that parents feel more protective of children in the below-11 age range compared with adolescents, where the response rate for community studies using a school-based ascertainment procedure in the UK is typically 50% (Goodyer et al., 2000).

All children had an estimated IQ above 80. The mean estimated IQ for the sample was 104 ($SD = 14$) and the mean age was 9 years, 5 months ($SD = 12$ months). The ethnic distribution in the sample was in line with regional statistics (Office of National Statistics, 2001) for eastern England (97% white, 2% of middle-eastern origin, 0.5% black and 0.5% Asian). Two procedures were employed to determine participation bias. First, permission was obtained for teachers of one of the schools to complete a screening measure of common emotional and behaviour problems, the Strengths and Difficulties Questionnaire (Goodman, 1997, 2001; Goodman, Ford, Simmons, Gatward, & Meltzer, 2000) on *all* the children in the school. Children who completed the SMFQ were compared with those who did not for their ratings on the SDQ. Independent sample *t*-tests revealed no evidence of any differences between the participants ($n = 61$) and non-participants ($n = 232$) when the five sub-scales of the SDQ (hyperactivity, emotional symptoms, conduct problems, peer problems, prosocial behaviour) were compared. Comparison of sociodemographic characteristics also revealed no difference between participants and non-participants.

Measures and Procedure

Short Mood and Feelings Questionnaire (Angold et al., 1995)

Our primary measure was the self-report SMFQ which comprises 13 items with a common response format: 0, never; 1, sometimes; 2, always. The SMFQ was administered individually at the same time as all the other measures. Due to concerns raised regarding the reading and understanding ability of younger children (Messer et al., 1995; Thapar & McGuffin, 1998), teachers were consulted as to the level of understanding for the 7-year-olds (youngest cohort), and it was decided that questions would be read aloud to this group (8%). As in previous studies that have used the SMFQ with younger children (Angold et al., 1995), the answers recorded were the participants' self-reports and not the examiners' opinions about them. Children in higher grades were invited to ask for help, if needed. However, none of the children in the high grades did so.

IQ

A shortened version (Vocabulary and Block Design subtests only) of the Wechsler Intelligence Scale for Children III (Wechsler, 1992), was used to estimate overall IQ in the sample. Sattler's (1988) guidelines were used for administration and scoring.

Psychopathology

To assess response bias, teachers completed the Strengths and Difficulties Questionnaire (SDQ; Goodman, 1997, 2001; Goodman et al., 2000). The SDQ was specifically designed to screen for psychiatric disorders in community samples and was shown to identify individuals with psychiatric diagnosis with a specificity (the proportion of people without disease who have a negative test result) of 94.6% (95% CI 94.1–95.1%) and a sensitivity of 63.3% (59.7–66.9%) (Goodman et al., 2000). Sensitivity (the proportion of people with disease who have a positive test result) for the SDQ has been demonstrated to be especially good for (70–90%) for identification of conduct, oppositional disorders and hyperactivity disorders.

Data Analytic Strategy

Combining CDFA with IRT

Our primary data analytic strategy was to apply two types of latent variable measurement models to the data: Categorical data factor analysis and item response (latent trait) theory (IRT). In the literature, applications of CDFA methodology often report only numerical results, whereas applications of related IRT models summarise their findings graphically. We wished to exploit both representations and therefore applied software for estimating both types of models.

Background to Statistical Models

Introductory, technical and statistical accounts of the family of the models that comprise the IRT and CDFA approach are presented elsewhere (Baker, 2001; Duncan-Jones, Grayson, & Moran, 1986; Lord, 1980; Lord & Novick, 1968; van der Linden & Hambleton, 1997). Bartholomew and Knott (1999) and McDonald (1999) provide excellent accounts of relations between these models and more traditional models.

Briefly, CDFA is an extension of the linear factor analysis model intended for dimension reduction and scaling of multiple binary or ordinal items. We consider only the probit-probit factor model of Muthén (1984, 1989). The model provides regression estimates of factor scores which are location values for latent trait scores along the single dimension of depression severity.

IRT analysis allows for detailed examination of the properties of individual items by determining item characteristic curves. ICCs characterize the interaction of a person with an item and plot the probability of a response (endorsing a symptom) given the level of the underlying characteristic measured by the "test" (Cooke & Michie, 1997)—in this case the severity of self-reported depressive symptoms. ICCs are defined in terms of two parameters that govern the shape and position of the S-shaped curves. In educational testing settings where IRT models were developed, the first parameter, a, is referred to as the item "discrimination" and governs the steepness of the slope of the ICC at the inflexion point of the S-shaped curve. A lower a-value is associated with a more gradual slope. Items with low slope estimates provide information over a wider range of latent trait values. Items with higher a-values have steeper curves and therefore discriminate more finely, but over a smaller range of latent trait values. The second parameter, b, relates to the prevalence of the item (here the proportion of the sample that endorse each SMFQ symptom). In educational settings, it is referred to as the "difficulty" parameter since there it is related to the difficulty of the test item. The b-parameter is related to the point on the trait dimension at which a respondent has a 50% probability of endorsing the item. In clinical and epidemiological studies, this parameter is referred to

as "commonality," since it relates to the prevalence of the symptoms. If the *b*-value is low, it indicates that the item/symptom is frequently endorsed even among low-trait individuals. In contrast, high values indicate that the item/symptom is likely to be endorsed only among high-trait individuals (e.g. severely depressed).

Graphical Representation of Scale Performance from IRT

The test information function (TIF) is a particularly useful output of an IRT analysis. The TIF profiles variations in the precision of measurement of the latent trait scores over the full range of estimated values (latent trait scores). The TIF provides a compelling graphical assessment of the effective measurement range of the SMFQ instrument. Reports of CDFA analyses do not usually provide any information on the range over which reliable (accurate) scores are estimated for individuals, only reliability at an aggregate level (for a sample group).

Software for Model Estimation

For the CDFA, we implemented Muthen's robust weighted least squares estimation approach using Mplus software (see Flora & Curran, 2004). For the IRT analyses, we implemented marginal maximum likelihood (MML) estimation of the two parameter logit-probit IRT model (Albanese & Knott, 1990; Bartholomew, 1987; Bartholomew & Knott, 1999; Bartholomew, Steele, Moustaki, & Gabraith, 2002) in TWOMISS. TWOMISS estimates model parameters for the logit-probit model by a maximum likelihood procedure using a modified expectation-maximisation (EM) algorithm.

MIMIC Modelling to Test for Differences in Item CDFA Intercepts (thresholds) by Age

In order to explore the impact of age on item responses, we used the MIMIC modelling approach of Gallo et al. (1994). This extends the categorical data factor model to include a covariate. MIMIC modelling enabled us to test the hypotheses of a linear association of latent trait scores with age (as a continuous measure), but no impact of age on the SMFQ item intercepts (thresholds). The presence of the latter would indicate differential item function with respect to age. We tested DIF for each SMFQ item (one at a time) by freely estimating the correlation of the latent factor with age, and also estimating the impact of age on the location of the threshold term in the item response model.

RESULTS

Descriptive Statistics

All participants who completed the SMFQ answered all questions because questionnaires were individually administered and checked before they were returned. There were 10 children who did not want to complete the questionnaire. Children were not required to give a reason for non-participation. Full questionnaire data were therefore available for 649 children. There were no missing data.

The SMFQ is a 13-item measure to which each response can be 2, 1 or 0. We recoded all responses to binary format (0/1) because the low frequency (<5%) of 2 scores meant that very little information would be lost in grouping these with 1 responses (options were recoded in the following way: $0 = 0$ and $1, 2 = 1$). Another advantage of recoding to binary responses (collapsing the two highest response categories) is that it provides a generally more parsimonious model, and simplifies the reporting of our IRT analysis of a candidate childhood depression screening tool.

A 2×2 comparison (7-year-olds versus older $\times$ score of 0 versus 1 on the SMFQ) on each item showed no significant differences between 7-year-olds (to whom questions were read out) and the rest of the children. Children 8 years and older did not display difficulty completing the questionnaires by themselves. In addition, age did not correlate with SMFQ total scores, although a significant relationship was found for SMFQ total scores and IQ ($r = -0.17$; $p < 0.01$).

Table II shows the prevalence of binary recoded item responses to each item: Five items were endorsed by 15% or more of the sample (item numbers: 4, restless; 7, poor concentration; 3, tired; 10, felt lonely; 12, never be as good); five items were endorsed by fewer than 10% (item numbers: 2, not enjoy anything; 6, cried a lot; 9, bad person; 11, unloved; 13, did everything wrong).

Confirmatory Factor Analysis in Mplus Using the Probit-Probit Model (Muthén, 1984, 1989)

First we summarize the results in terms of the lambda and tau parameters from the confirmatory categorical data factor analysis model. Then, we discuss indications of model fit [chi-square, root mean square error of

Table II. Item Endorsement in Terms of the Proportion of the Sample that Endorsed Each Individual Item, Tetrachoric Correlations Between Items, and Proportion of the Sample that Endorsed Pairs of Items

Item	%	1	2	3	4	5	6	7	8	9	10	11	12	13
1	0.107		0.023	0.043	0.045	0.036	0.028	0.049	0.036	0.028	0.049	0.032	0.043	0.028
2	0.094	0.331		0.036	0.029	0.031	0.022	0.046	0.032	0.011	0.034	0.017	0.037	0.020
3	0.176	0.412	0.356		0.065	0.040	0.037	0.063	0.042	0.020	0.049	0.039	0.056	0.039
4	0.275	0.233	0.060	0.181		0.054	0.036	0.091	0.059	0.029	0.060	0.045	0.062	0.042
5	0.114	0.474	0.448	0.335	0.315		0.031	0.049	0.054	0.025	0.059	0.039	0.062	0.032
6	0.090	0.439	0.362	0.403	0.188	0.468		0.031	0.031	0.019	0.040	0.022	0.032	0.028
7	0.178	0.490	0.510	0.401	0.430	0.456	0.299		0.056	0.036	0.065	0.042	0.068	0.048
8	0.114	0.474	0.473	0.357	0.373	0.686	0.468	0.531		0.028	0.051	0.032	0.054	0.036
9	0.060	0.589	0.209	0.258	0.286	0.507	0.447	0.565	0.567		0.031	0.023	0.028	0.019
10	0.164	0.523	0.361	0.284	0.175	0.599	0.478	0.455	0.509	0.509		0.048	0.057	0.031
11	0.088	0.525	0.261	0.435	0.341	0.594	0.386	0.478	0.500	0.557	0.593		0.043	0.023
12	0.151	0.474	0.440	0.037	0.239	0.662	0.388	0.526	0.578	0.477	0.449	0.559		0.043
13	0.077	0.495	0.380	0.497	0.370	0.551	0.554	0.628	0.600	0.493	0.397	0.470	0.618	

Note. Column 1 (item) refers to the 13 SMFQ items. Column 2 (%) refers to proportion of the sample that endorsed each item (responded 1or 2 rather than 0). The *lower left triangle* denotes the tetrachoric correlations between items. The *upper right triangle* denotes the proportion of the sample that endorsed both items. Item labels, 1: Miserable/unhappy; 2: Not enjoy anything; 3: Tired; 4: Restless; 5: No good anymore; 6: Cried a lot; 7: Poor concentration; 8: Hated self; 9: Bad person; 10: Felt lonely; 11: Unloved; 12: Never as good; 13: Everything wrong.

approximation (RMSEA) and standardised root mean square residuals (SRMR)].

Inspection of the factor loadings (column 3—lamda values) for the full sample shows high values for all items. Items 2, 3, 4 and 6 had the lowest factor loadings. The variance in item responses not explained by the single latent factor is quantified by the residual variances in column 4. Clearly these are quite substantial in magnitude for items 2, 3, 4 and 6. The standard errors (not shown) for these items were also larger than for other items. Overall the magnitude of most of the loadings is consistent with the hypothesis underpinning our use of a single latent variable, that is, that the SMFQ items are relatively sensitive (discriminating) indicators of an underlying continuum of depression.

Tests of Model Fit

All indices of model fit (chi-square, RMSEA and SRMR) supported the adequacy of a single latent variable model. The comparative fit index (CFI) (0.992) and Tucker Lewis Index (TLI) (0.994) were both high and close to 1; the root mean square error of approximation was low (<0.06) (RMSEA = 0.018). The standardized root mean square residual (0.063); the weighted root mean square residual were also low (0.833). There was little scope for model fit improvement by increasing the number of factors, and little justification in terms of our study aims (full results available on request from first author).

Full Information Item Factor Analysis Using the Logit-Probit Model (Bartholomew & Knott, 1999)

Pattern Frequencies

We first considered pattern frequencies for all response patterns for the binary recoded data. Under the full information approach, all information from the multiway cross-tabulation of all item responses is used. Thus, there are 13^2 possible response patterns for which frequencies may be reported. In most samples with more than six items, many possible patterns will not be observed (since sample size is usually much less than the number of items$^{\text{number of response categories}}$). Only with shorter instruments is it possible to display *all* pattern frequencies (i.e. for scales with 5 or 6 binary items). For a 13-item questionnaire and a sample of $n = 659$ like the current study, 10 pages of response patterns were returned. Space does not permit a full report of *all* response frequencies. Thus, only the most common patterns will be commented upon here. A full listing of response frequencies is available from the first author (CS). The three most common responses patterns are reported here.

Unsurprisingly, given the nature of the sample, the most common pattern was the absence of all morbidity [0000000000000] with a frequency of $n = 267$ (41%). This represents the modal pattern of children in the sample who had none of the 13 indicators of depression.

The second most common pattern was [0100000000000] with a frequency of 63 (10%)

indicating endorsement of only one item, item 2 (did not enjoy anything). The only other two-digit pattern frequency was 23 (3%) for the pattern [0001110000000] indicating endorsement of items 4 (restless), 5 (no good anymore) and 6 (cried a lot). This is the modal response for those children in the sample who had any symptoms (at least one item endorsed).

Parameter Estimates

Table III reports parameter estimates from TWOMISS for the maximum likelihood factor analysis under the Bartholomew and Knott (logit-probit) model. Here, the IRT factor analysis model is parameterised differently, using alpha 0 for item intercepts and alpha 1 for item slopes, but gives rise to very similar S-shapes for the item characteristic curves (see later).

The final column of Table III reports a transformation of the alpha 0 parameter showing the probability of endorsing an item for an individual at the median value on the latent trait. Here, these values range from lower to upper limits. So far we have focussed on the numerical results of the IRT and CDFA models. Next we consider the graphical representations that further facilitate the interpretation of these parameter estimates.

Table III. Corresponding Categorical Data Factor Analysis Results from Bartholomew and Knott's (1999) Logit-Probit Item Response Function Model

Item	Commonality α_{i0} (*SD*)	Discrimination α_{i1} (*SD*)	Symptom endorsement probability for median individual, π_{i0}
1. Miserable	3.12 (0.29)	1.80 (0.28)	0.04
2. Not enjoy	2.82 (0.231)	1.25 (0.23)	0.05
3. Tired	1.88 (0.152)	1.10 (0.17)	0.13
4. Restless	1.08 (0.105)	0.75 (0.13)	0.25
5. Felt no good	3.68 (0.432)	2.49 (0.40)	0.02
6. Cried a lot	3.09 (0.274)	1.49 (0.26)	0.04
7. Poor concentration	2.42 (0.246)	1.93 (0.27)	0.08
8. Hated self	3.64 (0.441)	2.44 (0.41)	0.02
9. Bad person	4.30 (0.541)	2.15 (0.43)	0.01
10. Felt lonely	2.39 (0.223)	1.71 (0.24)	0.08
11. Unloved	3.80 (0.476)	2.19 (0.42)	0.02
12. Never be as good	2.79 (0.275)	2.05 (0.28)	0.05
13. Everything wrong	4.28 (0.545)	2.44 (0.46)	0.01

Note. Under the logit-probit model, the alpha$_{i0}$ (α_{i0}) parameter in column 2 relates to the location on the *x*-axis of the inflexion point of the item response function curve. Alpha$_{i1}$ (α_{i1}) relates to the steepness of the slope of each ICC. Column 3 (π_{i0}), is a transformation of the α_{i0} parameter which denotes the probability of endorsing each SMFQ item for an individual with a latent trait score of zero, i.e. at the median/mean/modal value on the latent trait. The parameters have been estimated using full information marginal maximum likelihood estimation (modified EM algorithm) in TWOMISS software.

Graphic Representation of Test and Item Performance

Test Information Function (TIF)

We first report a graphic representation of the psychometric performance of the test as a whole (Fig. 1)—the TIF. The TIF is derived from the inverse of the posterior standard deviation of the latent trait estimates, when the factor scores (or latent traits estimates) are estimated using Bayesian (EAP) estimation. The TIF plots a function of the standard error. Taking the reciprocal of variance or standard deviation provides a humped plot with higher values indicating regions of precise measurement (small standard errors, relative to other regions)—this is "test information." As expected, the curve dips sharply at the end points where the SMFQ items provide little information. Figure 1 clearly shows that the most information (and therefore highest precision of measurement) is provided by the SMFQ around 1.5 standard deviations above the mean (0) on the latent trait.

Item Characteristic Curves (ICCs)

In Fig. 2, we represent the model results in terms of ICCs or "tracelines" in which the impact of the two IRT parameters are easily visible in relation to each other for each individual item. ICCs are S-shaped functions, plotted as a function of latent trait depression scores.

All items function at more or less the same level on the latent trait. Importantly, all items are located towards the more severe end—to the right of the figures. Both Figs. 1 and 2 suggest that a child located between 1 and 2 standard deviations above (worse) the population mean on the latent trait would have a 50% probability of endorsing the SMFQ items. We can also see that for a child at the mean (or median) latent trait value (0) the probability of endorsing any item is very low. This is also reflected in the entries in the last column of Table IV.

Although items 3 (tired) and 4 (restless) (top right two ICCs in Fig. 2) also function at the severe end of the latent trait, they show more shallow slopes. This indicates lower discriminating power with respect to the latent trait

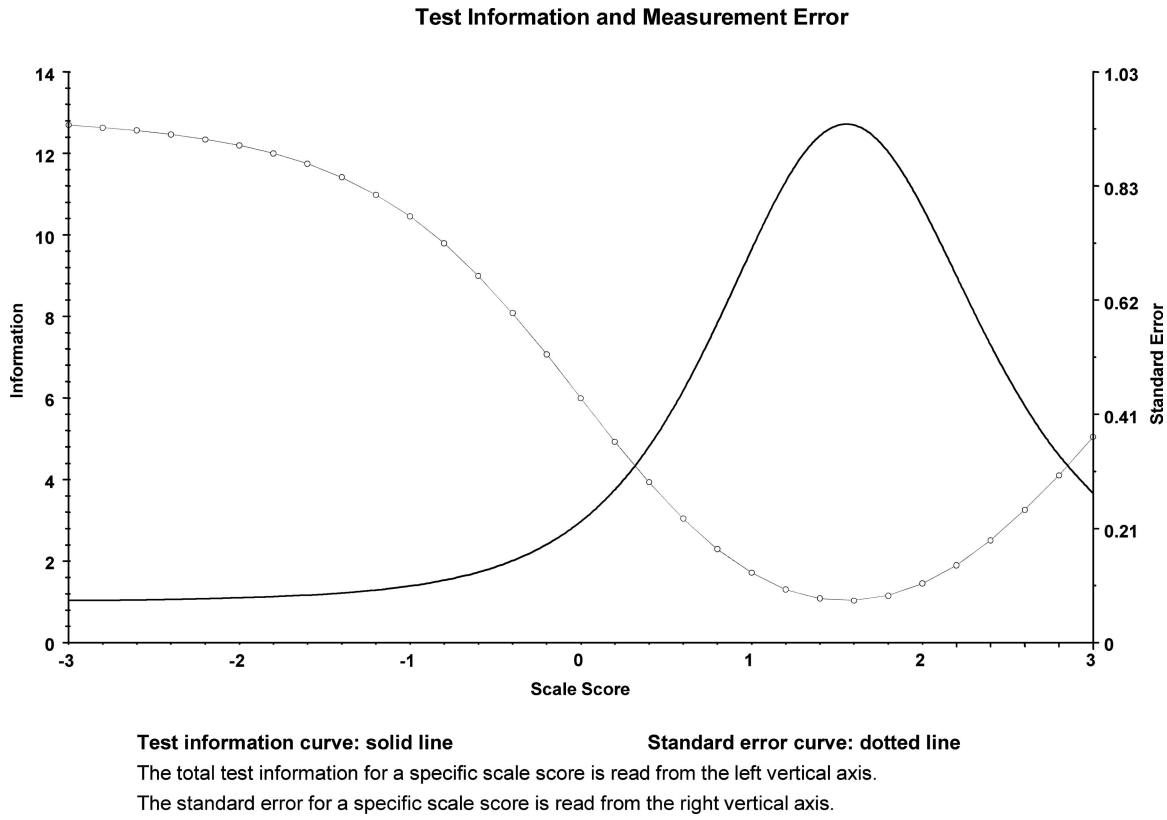

Fig. 1. Test characteristic curve and test information function for the SMFQ in a community sample of 7- through 11-year-olds.

(depression). These are thus the least sensitive items for measuring depression.

Items showing the highest commonality/lowest prevalence parameters were items 9: bad person ($\alpha_{i0} = -4.3058$) and 13: did everything wrong ($\alpha_{i1} = -4.2836$). Items 5: felt no good; 8: hated self and 13 showed the highest discrimination parameters ($\alpha_{i1} = 2.4916$; $\alpha_{i1} = 2.4440$ and $\alpha_{i1} = 2.4444$, respectively).

The Effect of Age

We examined the possibility of item bias (DIF) using an extension to IRT modelling referred to as MIMIC modelling (see data analytic strategy). Results showed that the ratio of estimate to standard error was <1.96 for the direct effect of the covariate on all items, indicating no significant item bias effects.

DISCUSSION

The current study is the first investigation in which a combined IRT/CDFA analysis, including MIMIC modelling, of the SMFQ has been carried out in 7–11-year-old children ascertained from the community. Methods for summarizing item responses were originally developed in the field of psychometrics and have been widely applied for the purposes of educational testing. They have found increasing application in medical (epidemiological) and health care (survey) research for the assessment of psychopathology, but have rarely been applied to evaluate clinical psychopathological assessments intended for use in children.

The advantages of IRT for abnormal psychology measurement are quite substantial (Cooke & Michie, 1997; Duncan-Jones et al., 1986; Embretson & Reise, 2000; Rouse et al., 1999; Santor et al., 1994). In scale construction and evaluation, IRT item analysis provides information that can be used for evaluating which items from a large pool should be used together to comprise a scale that is fit for that purpose (Waller, Tellegen, McDonald, & Lykken, 1996). IRT results can be used

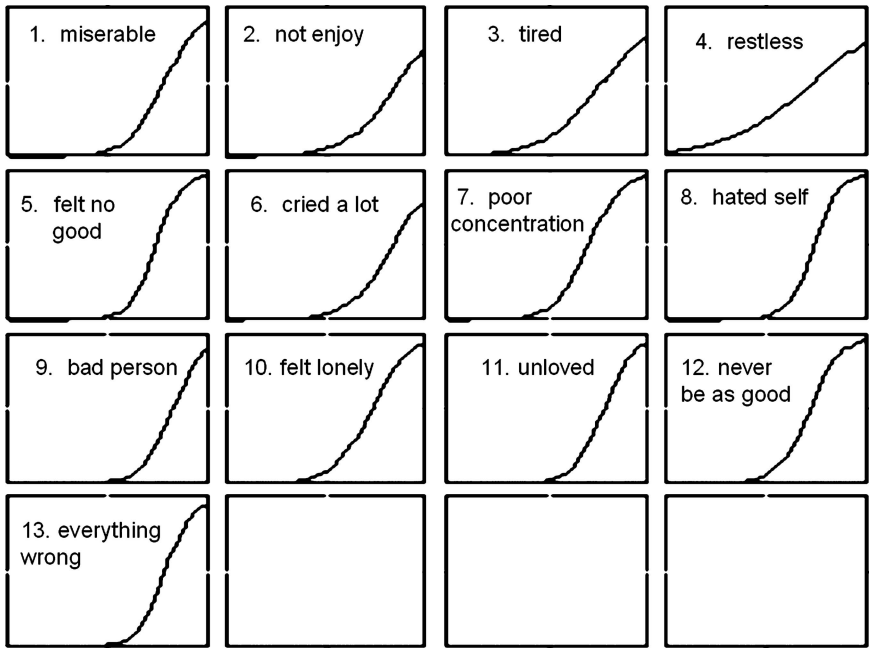

Fig. 2. Item characteristic curves for the 13 items of the SMFQ. Items 1–13 (panelled from *left* to *right*, first item *top left*). *Notes:* Axis labels removed for clarity of presentation. The *x*-axis is the estimated latent trait score which is distributed as a standard normal distribution; The *x*-axis ranges from -3 to $+3$. The *y*-axis is the probability that the SMFQ symptom is endorsed; the *y*-axis ranges from a minimum value of 0 to maximum value of 1.

Table IV. Confirmatory Factor Analyses using Mplus that Gives IRT Results for Categorical Response SMFQ Items (Binary Recoded)

	CDFA				IRT	
Item	Threshold (tau)	Loading (lambda)	Residual variance ($1 - \text{lambda}^2$)	*Z* test for loading (estimate/SE)	Discrimination parameter a	Intercept parameter b
1. Miserable	1.24	0.68	0.53	12.96	1.05	1.74
2. Not enjoy	1.31	0.54	0.70	8.19	0.76	2.22
3. Tired	0.93	0.53	0.71	9.21	0.68	1.66
4. Restless	0.59	0.40	0.83	6.79	0.48	1.35
5. Felt no good	1.20	0.79	0.36	19.74	1.42	1.50
6. Cried a lot	1.34	0.60	0.63	9.73	0.89	2.06
7. Poor concentration	0.92	0.72	0.48	15.16	1.13	1.26
8. Hated self	1.20	0.77	0.39	18.97	1.39	1.51
9. Bad person	1.55	0.70	0.50	12.43	1.22	2.04
10. Felt lonely	0.97	0.67	0.53	14.14	1.00	1.41
11. Unloved	1.35	0.72	0.47	15.07	1.26	1.76
12. Never be as good	1.03	0.74	0.44	17.45	1.18	1.37
13. Everything wrong	1.42	0.77	0.40	15.99	1.38	1.79

Notes. Parameters (loadings/lambda and thresnolds/tau) estimated using robust weighted least squares estimation (weighted lease squares estimation with mean– and variance-adjusted chi-square test statistic). Model fit under WLSMV: Chi-square value under robust WLS (mean and variance adjusted) $=$ 55.197, $df = 46$; $p = 0.16$; Comparative Fit Index $=$ 0.992; Tucker Lewis Index $=$ 0.994; Root mean square error of approximation $=$ 0.018; Weighted root mean square residual $=$ 0.833. Tau: Mplus thresholds; Lambda: Mplus factor loadings; IRT a-parameter: Discrimination parameter from response function parameterisation of CDFA, related to the factor loading (lambda); IRT Intercept (b-parameter): *x*-axis value where *p*(item) $=$ 0.5 in Fig. 2; CDFA denotes parameters from 'underlying variable parameterisation of the two parameter probit-probit IRT model; IRT denotes parameters under response function parameterisation of the two parameter probit-probit IRT model.

to estimate a person's trait level (latent trait score), which may be more accurate than summing unweighted individual item scores and can be calculated when there are partially missing data. IRT results are useful for determining whether the items included in a candidate scale exhibit differential item functioning for two populations (Cooke & Michie, 1997; Rouse et al., 1999). This particular advantage may be especially relevant in studies where developmental differences in the phenomenology of depression are the focus of interest (see Weiss & Garber, 2003 for a review of such studies). Lastly, using IRT technologies, there is potential for a clinical scale to be substantially shortened through the adoption of an adaptive testing approach (Gardner et al., 2004). With adaptive testing (computerised adaptive testing—CAT) a software program uses IRT parameters to select items in a sequence that optimally gain information on the severity score for each respondent, by tailoring questions to the likely level of the individual's score. In summary, latent trait modelling techniques (like IRT and CDFA) provide a closer approximation of the structural model underpinning psychiatric data than traditional linear methods (Cooke & Michie, 1997), thereby giving a better representation of empirical data and ultimately uncovering new aspects of a familiar instrument (Duncan-Jones et al., 1986).

By applying an IRT/CDFA approach we demonstrated, in line with Angold et al.'s findings (2002), that the SMFQ measures a unidimensional construct of depressive symptoms. Other studies have testified to the external validity of the SMFQ (see Table I). Based on these studies, and the face validity of SMFQ items, we assumed the latent trait underlying SMFQ items to be that of depression. However, in the absence of external validity data we cannot conclude definitively that the construct we studied was in fact depression. Notwithstanding this limitation, the current results support the SMFQ as a highly homogenous/unidimensional measure. Homogeneity/unidimensionality of a measure maximizes the ability to discriminate between diagnostically distinct groups (Costello & Angold, 1988) and thus speaks to the validity of the measure.

High correlations between SMFQ items in this sample suggest good internal consistency for the scale in younger children. Factor loadings for all SMFQ items were high, indicating that the items were highly discriminating with respect to the latent trait. Item distribution along the latent trait continuum was also consistent with the goals of the instrument to be used for screening purposes. All items were found to function at the severe end of the latent trait, with none being useful to measure individuals at the 50th percentile in the population (mean, median—the midpoint of the latent trait). According to these results, the SMFQ may not be appropriate for use in community studies where the interest lies in *variation* in average (mental) health, because this would require items to be more widely distributed *across* the range of the latent trait. As such, the SMFQ may be more appropriate for detecting children ages 7–11 who are likely to report high levels of depressive symptoms at the time of measurement. Despite the low response rate in the current study, which may affect the generalizability of our findings, the current study indicates with more appropriate statistical modelling techniques that the SMFQ does what it was designed to do (Angold, Costello, Pickles, & Winder, 1987).

An interesting next step would be to carry out IRT analyses of the SMFQ in a clinical sample with a current diagnosis of depression. This should yield a similar pattern of item loadings but the score distributions would be skewed toward the depressed end of the latent trait. In our sample, most children endorsed zero scores on the SMFQ, which is what is expected for community samples.

Our analyses showed that two items function a little differently than the others (3: tired and 4: restless). These items were characterised by ICCs that exhibited shallower slopes and leftmost thresholds. They therefore offer almost no discriminating power at the most severe end of the depressive latent trait but would contribute to the discrimination of individuals with lower scores. This does not, however, provide a strong case for shortening the SMFQ any further. Although these items exhibited shallower slopes and lower thresholds in comparison with other items, their psychometric properties are not sufficiently different from the other items to exclude them. In contrast, items 8: hated self and 13: did everything wrong showed most discriminating power, whilst items 9: bad person and 13: did everything wrong functioned at the most severe end of the latent trait. This suggest that these questions may be the most useful to ask if a clinician or researcher is interested in identifying a child with depression with expediency.

The information gained from these analyses regarding individual item performance is also of theoretical and substantive interest. Items 3: tired and 4: restless which did not load as highly (but still sufficiently to retain them) on to the latent trait can be seen as less central to the trait for children ages 7–11 compared with items 8: hated self, 9: bad person, 11: unloved, 12: never be as good and 13: did everything wrong that load highly. Within this age range, cognitive symptoms of depression such as items 8, 9, 11 and 12 show higher factor loadings, and therefore more discriminatory power at the severe end of the latent trait compared with "somatic" symptoms like items 3 and 4. These results are in line with previous research showing

that both cognitive and somatic symptoms in the long form of the MFQ are essential to the construct of depression in children and adolescents (Kent et al., 1997), but may differ in their prevalence and/or loading on a single factor of severity. The current results showed somatic symptoms of depression (tired and restless) contributed less than cognitive symptoms to the underlying unidimensional latent trait of depression in 7–11–year-old children.

In summary, the IRT analyses conducted here suggest that the SMFQ: (1) is a unidimensional or homogeneous measure; (2) has items showing relatively good discrimination at the severe end of the latent trait, indicating that the instrument, in terms of its internal construct validity, is an adequate screening measure for depressive symptoms in a community sample of children ages 7–11-years-old; (3) although this study lacked an adolescent comparison sample, these data suggest that the SMFQ is not biased with respect to age from 7- to 11-years-old and is suitable for use with children as young as 7.

ACKNOWLEDGMENTS

The authors are grateful to all the families and schools who participated. We also wish to thank Sarah Moore, Heather Brown, Penny Hazell and Maria Loades for helping with data collection and entry, and Dr. Melanie Merricks and Dr. Carol Stott for valuable discussion. Carla Sharp was supported by an NHS Post-Doctoral Fellowship, University of Cambridge, UK. Tim Croudace was supported by a Dept of Health Career Scientist Award (Public Health).

REFERENCES

Albanese, M. T., & Knott, M. (1990). *Twomiss: A computer program for fitting a one- or two- factor logit-probit latent variable model to binary data when observations may be missing*. London, UK: London School of Economics and Political Sciences.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental Disorders—Fourth Edition (DSM-IV)*. Washington: American Psychiatric Association.

Angold, A., Costello, E. J., Messer, S. C., Pickles, A., Winder, F., & Silver, D. (1995). Development of a short questionnaire for use in epidemiological studies of depression in children and adolescents. *International Journal of Methods in Psychiatric Research*, *5,* 237–249.

Angold, A., Costello, E. J., Pickles, A., & Winder, F. (1987). *The development of a questionnaire for use in epidemiological studies of depression in children and adolescents*. London: Medical Research Council Child Psychiatry Unit.

Angold, A., Erkanli, A., Silberg, J., Eaves, L., & Costello, E. J. (2002). Depression scale scores in 8–17-year-olds: Effects of age and gender. *Journal of Child Psychology and Psychiatry*, *43,* 1052–1063.

Baker, F. (2001). *The basics of item response theory* (2nd ed.). ERIC Clearinghouse on Assessment and Evaluation. Retrieved from http://edres.org/irt/

Bartholomew, D. J., Steele, F., Moustaki, I., & Gabraith, J. (2002). *The analysis and interpretation of multivariate data for social scientists*. London: Chapman & Hall/CRC.

Cheong, Y. F., & Raudenbush, S. W. (2000). Measurement and structural models for children's problem behaviors. *Psychological Methods*, *5,* 477–495.

Cooke, D. J., & Michie, C. (1997). An item response theory analysis of the Hare Psychopathy Checklist. *Psychological Assessment*, *9,* 3–14.

Costello, E. J., & Angold, A. (1988). Scales to assess child and adolescent depression: Checklists, screens and nets. *Journal of the American Academy of Child and Adolescent Psychiatry*, *27,* 726–737.

Costello, E. J., Angold, A., Burns, B. J., Erkanli, A., Stangl, D. K., & Tweed, D. L. (1996a). The great smoky mountains study of youth: Functional impairment and serious emotional disturbance. *Archives of General Psychiatry*, *53,* 1137–1143.

Costello, E. J., Angold, A., Burns, B. J., Stangl, D. K., Tweed, D. L., Erkanli, A., et al. (1996b). The great smoky mountains study of youth: Goals, design, methods, and the prevalence of DSM-III-R disorders. *Archives of General Psychiatry*, *53,* 1129–1136.

Duncan Jones, P., Grayson, D. A., & Moran, P. A. (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*. London: Lawrence Erlbaum.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9,* 466–491.

Gallo, J. J., Anthony, J. C., & Muthén, B. O. (1994). Age differences in the symptoms of depression: A latent trait analysis. *Journals of Gerontology: Psychological Sciences*, *49,* 251–264.

Gardner, W., Shear, K., Kelleher, K. J., Pajer, K. A., Mammen, O., Buysse, D., et al. (2004). Computerized adaptive measurement of depression: A simulation study. *BMC Psychiatry*, *4,* 13.

Goodman, R. (1997). The strengths and difficulties questionnaire: A research note. *Journal of Child Psychology and Psychiatry*, *38,* 581–586.

Goodman, R. (2001). Psychometric properties of the strengths and difficulties questionnaire. *Journal of the American Academy of Child and Adolescent Psychiatry*, *40,* 1337–1345.

Goodman, R., Ford, T., Simmons, H., Gatward, R., & Meltzer, H. (2000). Using the strengths and difficulties questionnaire (SDQ) to screen for child psychiatric disorders in a community sample. *British Journal of Psychiatry*, *177,* 534–539.

Goodyer, I. M., Herbert, J., Tamplin, A., & Altham, P. M. (2000). First-episode major depression in adolescents: Affective, cognitive and endocrine characteristics of risk status and predictors of onset. *British Journal of Psychiatry*, *176,* 142–149.

Goodyer, I. M., & Sharp, C. (2005). Childhood depression. In B. Hopkins, R. G. Barr, G. F. Michel, & P. Rochat (Eds.), *The cambridge encyclopedia of child development* (pp. 420–423). Cambridge: Cambridge University Press.

Hankin, B. L., Abramson, L. Y., Moffitt, T. E., Silva, P. A., McGee, R., & Angell, K. E. (1998). Development of depression from preadolescence to young adulthood: Emerging gender differences in a 10-year longitudinal study. *Journal of Abnormal Psychology*, *107,* 128–140.

Kent, L., Vostanis, P., & Feehan, C. (1997). Detection of major and minor depression in children and adolescents: Evaluation of the Mood and Feelings Questionnaire. *Journal of Child Psychology and Psychiatry*, *38,* 565–573.

Lambert, M. C., Schmitt, N., Samms-Vaughan, M. E., An, J. S., Fairclough, M., & Nutter, C. A. (2003). Is it prudent to administer all items for each Child Behavior Checklist cross-informant syndrome? Evaluating the psychometric properties of the youth self-report dimensions with confirmatory factor analysis and item response theory. *Psychological Assessment*, *15,* 550–568.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Mahwah, NJ: Erlbaum.

Lord, F. M., & Novick, M. R. (Eds.). (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: LEA.

Messer, S. C., Angold, A., Costello, E. J., Loeber, R., Van Kammen, W., & Stouthamer-Loeber, M. (1995). Development of a short questionnaire for use in epidemiological studies of depression in children and adolescents: Factor composition and structure across development. *International Journal of Methods in Psychiatric Research*, *5,* 251–262.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49,* 115–132.

Muthén, B. (1989). Multiple-group structural modelling with non-normal continuous variables. *British Journal of Mathematical and Statistical Psychology*, *42,* 55–62.

Office of National Statistics. (2001). Retrieved from http://www.statistics.gov.uk/

Park, R. J., Goodyer, I. M., & Teasdale, J. D. (2002). Categoric overgeneral autobiographical memory in adolescents with major depressive disorder. *Psychological Medicine*, *32,* 267–276.

Patton, G. C., Carlin, J. B., Shao, Q., Hibbert, M. E., & Bowes, G. (1997). Adolescent dieting: Healthy weight control or borderline eating disorder? *Journal of Child Psychology and Psychiatry*, *38,* 299–306.

Rouse, S. V., Finger, M. S., & Butcher, J. N. (1999). Advances in clinical personality measurement: An item response theory analysis of the MMPI-2 PSY-5 scales. *Journal of Personality Assessment*, *72,* 282–307.

Santor, D. A., Ramsay, J. O., & Zuroff, D. C. (1994). Nonparametric item analyses of the Beck Depression Inventory: Evaluating gender item bias and response option weights. *Psychological Assessment*, *6,* 255–270.

Sattler, J. M. (1988). *Assessment of children*. San Diego, CA: Sattler.

Shrout, P. E., & Parides, M. (1992). Conventional factor analysis as an approximation to latent trait models for dichotomous data. *International Journal of Methods in Psychiatric Research*, *2,* 55–65.

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

Thapar, A., & McGuffin, P. (1998). Validity of the shortened Mood and Feelings Questionnaire in a community sample of children and adolescents: A preliminary research note. *Psychiatry Research*, *81,* 259–268.

van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York: Springer-Verlag.

Waller, N. G., Tellegen, A., McDonald, R. P., & Lykken, D. T. (1996). Exploring nonlinear models in personality assessment: Development and preliminary validation of a negative emotionality scale. *Journal of Personality*, *64,* 545–576.

Wechsler, D. (1992). *Wechsler intelligence scale for children* (3rd UK ed.). London: Psychological Corporation.

Weiss, B., & Garber, J. (2003). Developmental differences in the phenomenology of depression. *Development and Psychopathology*, *15,* 403–430.

Wood, A., Kroll, L., Moore, A., & Harrington, R. (1995). Properties of the Mood and Feelings Questionnaire in adolescent psychiatric outpatients: A research note. *Journal of Child Psychology and Psychiatry*, *36,* 327–334.