British Journal of Clinical Psychology (2006), 45, 185–203
© 2006 The British Psychological Society

The British Psychological Society

www.bpsjournals.co.uk

# Dimensions of variation on the CORE-OM

K. Jake Lyne[1]*, Paul Barrett[2], Chris Evans[3] and Michael Barkham[4]

[1]Department of Psychology, University of York and Selby and York Primary Care Trust, UK
[2]Department of Management and Employment Relations, University of Auckland, New Zealand
[3]Rampton Hospital, Nottinghamshire Healthcare NHS Trust, UK
[4]Psychological Therapies Research Centre, University of Leeds, UK

**Background.** The Clinical Outcomes in Routine Evaluation-Outcome Measure (CORE-OM) is a self-report measure comprising 28 items tapping three domains; subjective well-being, psychological problems and functioning. In addition to the potential theoretical value of the domains for operationalizing the phase model of psychotherapy, when consulted, managers and clinicians considered the distinction between problems and functioning important for assessing case-mix and clinical outcomes. A further domain comprising six items was included to indicate possible risk. Subsequent analysis has suggested an alternative structure for CORE-OM with factors for risk and positively and negatively worded items (Evans *et al.*, 2002).

**Methods.** This study compares models for the interpersonal factor structure in data from the CORE-OM in 2,140 patients receiving psychological therapy in the UK.

**Results.** A multi-method, multi-trait, nested factors solution accounted optimally for the CORE-OM item covariance, with a first-order general factor latent and residualized first-order factors of subjective well-being, psychological problems, functioning and risk and with positively and negatively worded methods factors. The general factor was labelled psychological distress. Scale quality for CORE-OM, using a scoring method in which non-risk items are treated as a single scale and risk items as a second scale is satisfactory.

**Implications.** The CORE-OM has a complex factor structure and may be best scored as 2 scales for risk and psychological distress. The distinct measurement of psychological problems and functioning is problematic, partly because many patients receiving out-patient psychological therapies and counselling services function relatively well in comparison with patients receiving general psychiatric services. In addition, a clear distinction between self-report scales for these variables is overshadowed by their common variance with a general factor for psychological distress. An alternative strategy for operationalizing this distinction is proposed.

*Correspondence should be addressed to Dr Jake Lyne, Department of Psychological Therapies, The Old Chapel, Bootham Park, York YO30 7BY, UK (e-mail: jake.lyne@sypct.nhs.uk).
*This paper, which is jointly authored by the current editor, was processed and accepted for publication by the previous editorial team.*

DOI:10.1348/014466505X39106

In 1970, an invited workshop discussed the possibility of developing a core outcome battery for use across studies focusing on the evaluation of the psychological therapies (Waskow, 1975). This was revisited in the mid-1990s in an American Psychological Association (APA) scientific meeting (Strupp, Horowitz, & Lambert, 1997). In parallel, a UK initiative moved forward the vision of a core outcome battery via a specifically designed measure (Barkham *et al.*, 1998), which would be suitable for routine use and tap core components of presenting problems across the widest range of clients. The resultant measure was named the Clinical Outcomes in Routine Evaluation-Outcome Measure (CORE-OM).

The CORE-OM is a 34-item self-report questionnaire designed for use as a baseline and outcome measure in psychological therapies (Evans *et al.*, 2002, 2000). It comprises high and low intensity items tapping domains for *subjective well-being* (4 items), *problems* (12 items) and *functioning* (12 items). There are also six items to measure *risk*, four for *risk of self-harm* and two for *risk of harm-to-others*. A programme of work has set out the rationale (Barkham *et al.*, 1998), development (Evans *et al.*, 2000), psychometric properties (Evans *et al.*, 2002), and clinical utility (Barkham *et al.*, 2001) of the CORE-OM. The measure was developed after consultation with service providers and purchasers, who placed high priority on the measurement of symptoms at intake, and reduction in symptoms and improvement in functioning as a result of therapy or counselling (Barkham *et al.*, 1998). The three non-risk domains were considered to be compatible with the phase model of change, which suggests a sequential impact on phase domains of subjective well-being early in therapy, progressing to symptoms and then to aspects of life functioning (Howard, Lueger, Maling, & Marinovic, 1993).

The selection and design of the items was intended to be acceptable to patients, and to be comprehensible as one core part of any routine nomothetic evaluation system for therapists working from varied theoretical schools. The measure is copyleft, which encourages non-profit organizations to make copies provided the measure is not changed. The CORE-OM is part of an information management system termed the CORE System (Mellor-Clark, Barkham, Connell, & Evans, 1999). This consists of therapist-completed baseline and outcome evaluation *pro forma* comprising data on client demographics, case-mix, and service trajectory (Mellor-Clark & Barkham 2000).

Initial results on the CORE-OM from large clinical datasets and non-clinical samples of convenience showed large differences between the clinical and non-clinical samples (Evans *et al.*, 2002). The CORE-OM has shown large and statistically significant within-patient change over therapy, and has been used to benchmark services revealing small but significant differences between patients on baseline, outcome, and change scores (Barkham *et al.*, 2001). Subsequent uptake of the measure has been extensive within the UK and is not restricted to any particular theoretical persuasion of service or therapist.

The analyses of the psychometric properties of the CORE-OM (Evans *et al.*, 2002) explored six scores from the 34 items: the mean item totals for each of the four domains, the mean item total across the 28 non-risk items, (reflecting a possible hierarchical relationship between the three non-risk domains and a general measure of psychological distress), and the mean item total for all 34 items. The results from 1,106 non-clinical and 890 clinical participants showed strong positive correlations in the clinical sample between these scores (e.g. .77 between subjective well-being and problems), with the lowest correlations between risk and other scores (e.g. .33 between risk and subjective well-being). Internal reliabilities (.75–.94) and 1 week test–retest reliabilities (.60–.91)

were very acceptable. Principal component analyses showed a large proportion of the variance in the first component (e.g. 38% for the non-clinical sample).

In both clinical and non-clinical samples, oblique rotation of three components gave negatively keyed and positively keyed non-risk items, with a third component for risk items. The separation was perfect in the non-clinical sample but slightly compromised in the clinical sample. The differentiation into factors for positively and negatively worded items (NW) suggests an alternative model for CORE-OM with two method factors regressing on to a general factor latent for psychological distress.

Convergent correlations of the CORE-OM scores against various measures were explored in several clinical samples. These were: the four scales of the 28-item General Health Questionnaire (GHQ; Goldberg & Hillier, 1979), the original and second versions of the Beck Depression Inventory (BDI; Beck, Steer, & Brown, 1996; Beck, Ward, Mendelson, Mock, & Erbaugh, 1961), the Beck Anxiety Inventory (BAI; Beck, Epstein, Brown, & Steer, 1988), the Brief Symptom Inventory (BSI; Derogatis & Melisaratos, 1983), the revised version of the Symptom Checklist-90-R (Derogatis, 1983) and the 32-item version of the Inventory of Interpersonal Problems (IIP-32; Barkham, Hardy, & Startup, 1996; Horowitz, Rosenberg, Baer, Ureno, & Villasenor, 1988). There were strong, positive correlations of the referential measures with the CORE-OM scales, and some evidence for discriminant validity where the strongest correlation was with the scale that would have been expected (e.g. IIP-32 with functioning). However, for only one out of 11 comparisons was the effect statistically significant (BAI with problems, $N = 218$, $r = .68$; compared with BAI with functioning, $N = 218$, $r = .55$; correlation difference $p < .05$). Thus discriminant validity was not strong in relation to the huge general covariance between all of the measures. The CORE-OM risk items were found to have strong, positive correlations with a therapist rating of risk in a student counselling client sample.

Finally, although CORE-OM might have been expected to be useful for testing the phase model of psychotherapy, Shapiro *et al.* (2001) found no support for the phase model of change based on repeated measurement in psychotherapy patients using CORE-OM as the change measure. The pattern of strong correlations between the domains was repeated at each measurement point. Notwithstanding problems with the phase model (Joyce, Ogrodniczuk, Piper, & McCallum, 2002) the stability in correlations between the domains over time raises the question as to whether the CORE-OM domains covary too strongly to test it.

In summary, the initial work on the CORE-OM supports its utility as a pragmatic measure, most strongly supporting separation of the risk items from the remaining 28 items. However, differentiation between the three domains (subjective well-being, problems, and functioning) remains to be determined. The purpose of this study is to report a detailed exploration of CORE-OM by comparing models for its psychometric structure to determine which is optimal. It also assesses whether the four self-harm risk items have psychometric claims to form a scale rather than be used only as clinical flags.

## Method

### *Data set*

The study used completed questionnaires from clients referred to counselling and psychological therapy services in the UK.

Of 2,277 cases, data were missing for gender in eight cases. In the remaining 2,269 cases, the distribution of missing data was 1,919 cases (84%) with no missing data, 221

cases (10%) with one item of missing data and 129 cases (6%) with more than one item of missing data. The latter were rejected, leaving 2,140 cases. Missing data were replaced using means substitution in the 221 cases that had one item of missing data.

The 2,140 cases were subsequently divided into subsets for men and women (males, $N = 590$; females, $N = 1,550$) in order to test whether the factor analyses would hold across the sexes. The mean age for men was 35.49 years ($SD = 13.77$) and for women 34.46 years ($SD = 13.55$). Marital and ethnic data were not collected.

### *Analytical methods*

#### *Gender matrix comparison*

As indicated above, subsets for men and women were formed to test whether the CORE-OM item correlations might be considered homogeneous across gender. Given that this analysis was concerned with testing a hypothesis of gender correlation matrix similarity prior to structural modelling, the analysis technique selected was that proposed by Steiger (1980a, 1980b) whereby a matrix of correlations may be compared with another matrix using the same variables. This technique provides a chi-squared significance test and other close-fit tests of matrix homogeneity. This was achieved using the STATISTICA SEPATH Structural Equation Modelling software (StatSoft Inc., 2003).

#### *Score correlations*

The published score key was used to derive scores from the CORE-OM for well-being, psychological problems, functioning, and risk for 2,140 cases. Item Likert scores (range 0–4) were summed to give four domain scores for each respondent, and a Pearson product moment correlation matrix was computed for the four domain scores.

#### *Scale quality*

In view of the strong association between the CORE-OM domains, scale quality was explored by calculating internal reliability using coefficient alpha (Cronbach, 1951) and confidence intervals using the methods proposed by Feldt, Woodruff, and Salhi (1987). Item complexity analysis was used to assess the quality of each of the published CORE-OM domain scores. The full details of the procedures have been reported elsewhere (Barrett, Kline, Paltiel, & Eysenck, 1996). Essentially the analyses identified three parameters: the internal reliability of scales, the item complexity (i.e. the number of items that correlate above a specified level with more than one scale), and the signal-to-noise ratio between the different scales (i.e. the extent to which any specific item correlates with its own scale in comparison with its correlations with other scales to which it is presumed not to belong). These parameters were combined in a formula to calculate an index of the measurement quality of scales (the Scale quality index), where 0 indicates *lowest possible quality* and 1 indicates *perfect quality*. Summary quality indexes were calculated for the published CORE-OM score key for two methods of scoring.

#### *Structural equation modelling (SEM) analyses*

Given the initial work and exploration of the CORE-OM reported above, it was considered appropriate to undertake a model-based assessment strategy investigating the measurement and structural model of the questionnaire. A selection of 10 models were examined for their fit to the covariance matrix data. Conditional upon the gender

matrix comparison results, these models would be fit to either separate gender matrices or the total sample matrix of 2,140 cases.

The first model tested was the established CORE-OM model and this was then systematically varied in terms of alternative conceptualizations of the CORE-OM structure. This model-testing process is designed as a comprehensive series of tests whose aim is to try to establish the optimum psychometric factor model for CORE-OM. All SEM analyses were conducted using both SPSS AMOS-5 (Arbuckle, 2003) and STATISTICA SEPATH software, to provide an internal check on modelling setup and specifications, as well as the calculation of the McDonald (1989) Non-centrality Index within SEPATH. SEM analyses were undertaken on covariance matrices in AMOS, with output expressed in standardized form (which corresponds to the completely standardized solution in SEPATH). Five indices were used to gauge goodness-of-fit of each model-implied covariance matrix to the sample covariance matrix. These were the chi-squared exact-fit test, and four close-fit tests comprising the McDonald non-centrality index (Mc), Steiger and Lind's (1980) root mean squared error of approximation (RMSEA), Bentler's (1990) comparative fit index (CFI), and the standardized root mean square residual index (SRMR). The chi-squared alpha for null-hypothesis testing was set at .05. Suggested minimum values indicating acceptable model-fit for the close-fit test indices were taken from Hu and Bentler (1999). These were .90 or greater for Mc, .06 or less for RMSEA, .95 or greater for CFI, and .08 or less for SRMR.

General factor models for CORE-OM were tested using the nested factors model (Gustafsson & Balke, 1993). This is a method of determining the extent to which each first-order latent variable can be considered causal for the covariation between a group of manifest variable responses; this covariation is modelled as independent from the causal influence of a hypothesized general first-order latent variable. In essence, each latent variable is 'residualized' with respect to the general factor latent (i.e. is orthogonal to it).

Difference in goodness-of-fit between models was tested using the chi-squared test.

## Results

### *Gender matrix comparison*

The results of the analysis comparing the two CORE-OM correlation matrices from male and female respondents for homogeneity of coefficients yielded a chi-squared goodness-of-fit statistic of 819.06, $df = 561$, $p < .0001$, which indicated an exact-fit rejection of the null hypothesis of no-difference. However, the close-fit index of RMSEA was .019, with 90% confidence interval between .016 and .022, and SRMR = .031. The largest standardized residual correlation was just .035. This value, along with the information from the close-fit tests, indicated that the two gender matrices might reasonably be considered homogeneous, given we are content to accept that the largest observed discrepancy between the residual correlations is as low as .035. The mean standardized residual correlation discrepancy is .001. Therefore, both male and female datasets were combined into a total sample dataset of $N = 2,140$, which was used in all subsequent analyses.

### *Domain score correlations and scale quality*

A Pearson's product moment correlation matrix was calculated for the sums of scores for each of the four CORE-OM domains for all 2,140 cases. All correlations in Table 1 are significant at $p < .0001$.

**Table 1.** Correlations between the CORE-OM subscales

| | CORE-OM domains | | |
|---|---|---|---|
| | Problems | Functioning | Risk |
| Subjective well-being | .79 | .75 | .46 |
| Problems | | .76 | .50 |
| Functioning | | | .53 |

Cronbach's alpha coefficients were calculated for each score (upper section of Table 2), as an estimate of internal consistency reliability.

All 2,140 cases were used for an assessment of scale quality for two scoring methods, the first for the four CORE-OM domains, and the second with two scales; the 28 non-risk items and the six risk items. The results are shown in Table 2. The scale quality for the whole test using the domain scoring method was .18 (this improves marginally to .24 if the two risk-to-others risk items 6 and 22 are removed from the risk subscale) confirming that the original score key provides scale scores which are substantially overlapping with each other, and therefore does not reveal separate dimensions of differences between clients. Scoring the test for a 28-item psychological distress scale and a 6-item risk scale results in a much improved overall scale quality score of .59, which rises to .60 if the two risk-to-others items are removed.

**Table 2.** Item analyses and assessments of scale quality for two methods of scoring CORE-OM

| Domains | Items | Mean | *SD* | Coefficient alpha 95% CI | Mean item-total correlation | Scale quality |
|---|---|---|---|---|---|---|
| Subjective well-being | 4 | 2.33 | 0.92 | .74 (.72–.76) | .54 | .00 |
| Problems | 12 | 2.23 | 0.83 | .87 (.86–.87) | .57 | .11 |
| Functioning | 12 | 1.75 | 0.78 | .85 (.84–.86) | .53 | .11 |
| Risk | 6 | 0.46 | 0.64 | .77 (.75–.79) | .55 | .48 |
| Psychological distress | 28 | 2.04 | 0.76 | .93 (.93–.94) | .57 | .57 |

### *Structural equation modelling*

The results of the model fitting exercises are presented in Table 3, and the models are shown in Figure 1.

#### *Tests of the established CORE-OM model*

Model 1a was the first to be examined. This corresponds to the established CORE-OM model, comprising four latent variables of subjective well-being, functioning, psychological problems and risk, where all latent variables are modelled as correlated (including the risk latent). As can be see from Table 3, this model did not fit very well to the data.

Examination of the standardized path coefficients for the risk latent to its manifest items, confirmed that the risk-to-others items 6 and 22 were poorly related to the risk latent. Excluding these items in Model 1b, whilst retaining the correlation between all latents as in Model 1a, produced a slight improvement in close-fit, although the fit

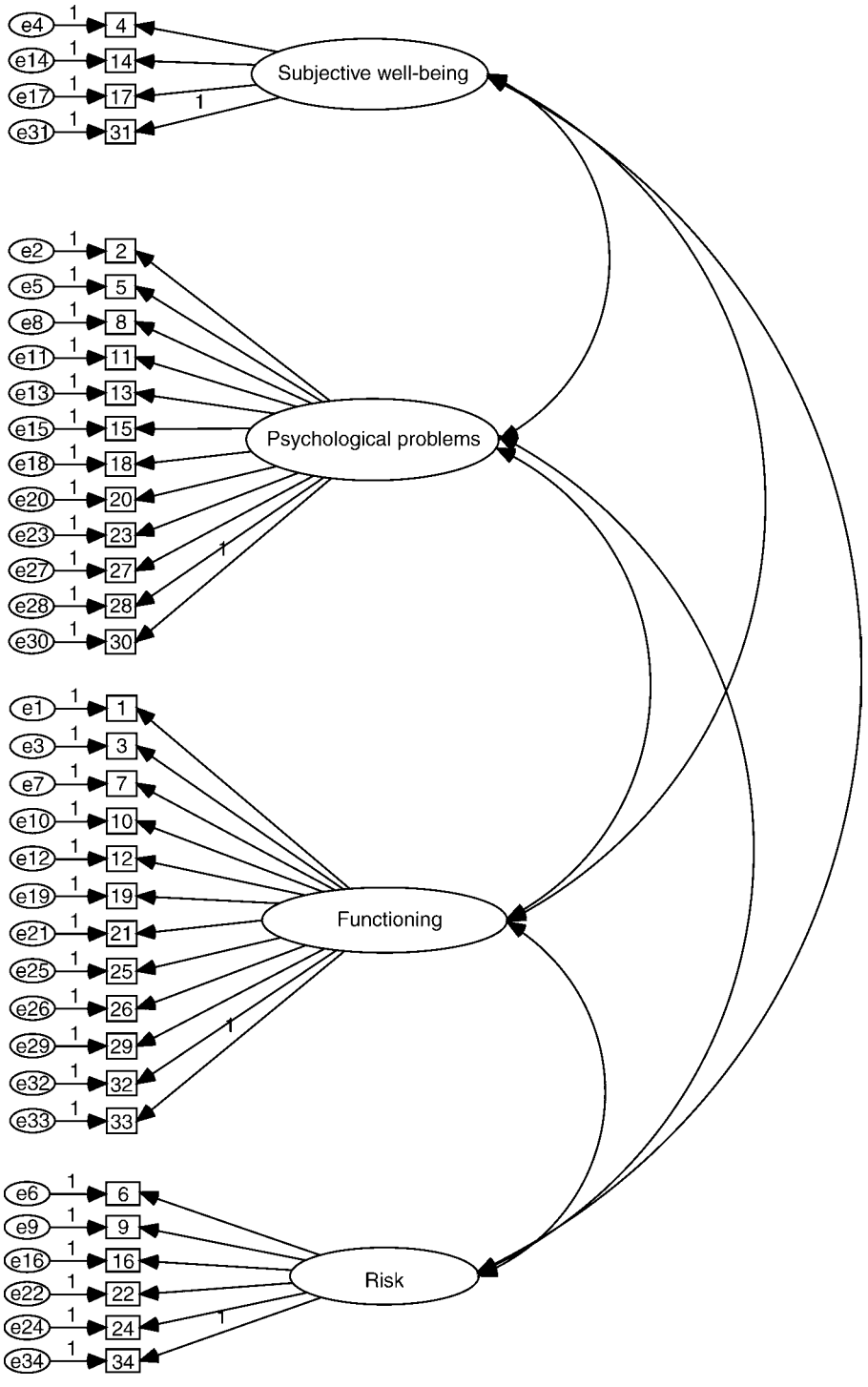


**Figure 1.** The structural equation models. Model 1a: 34-item, four correlated latent variables. For all subsequent models, variables 6 and 22 were removed. For Model 2 subjective well-being and psychological problems were combined. For Model 3a the first three latents were combined and for 3b all four latents were combined. (*Continued*).

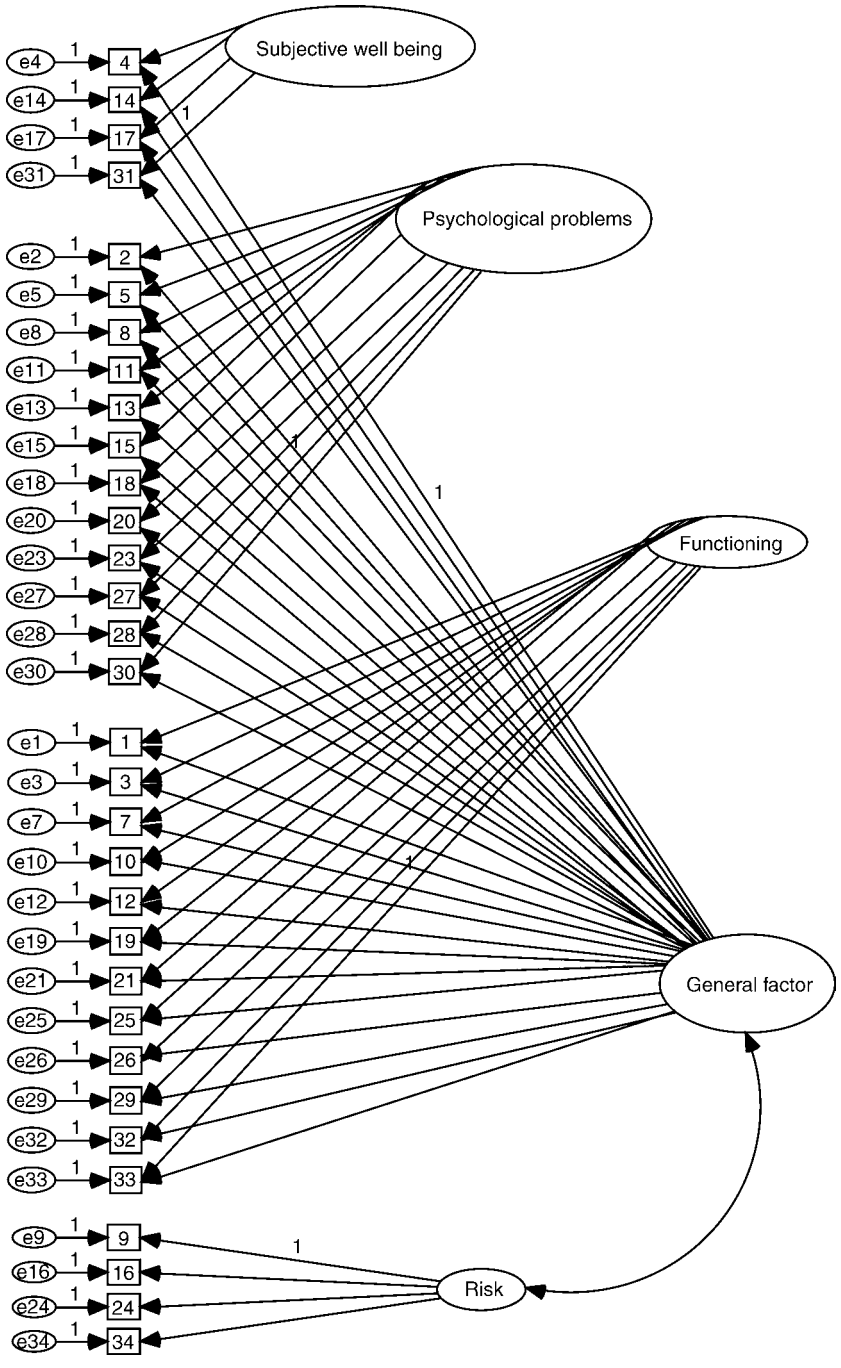


**Figure 1.** *(Continued)* Model 4a: a nested factors first-order general factor model with three residualized latents and risk as a correlated latent. For Model 4b risk was residualized (i.e. not correlated as shown here for Model 4a).

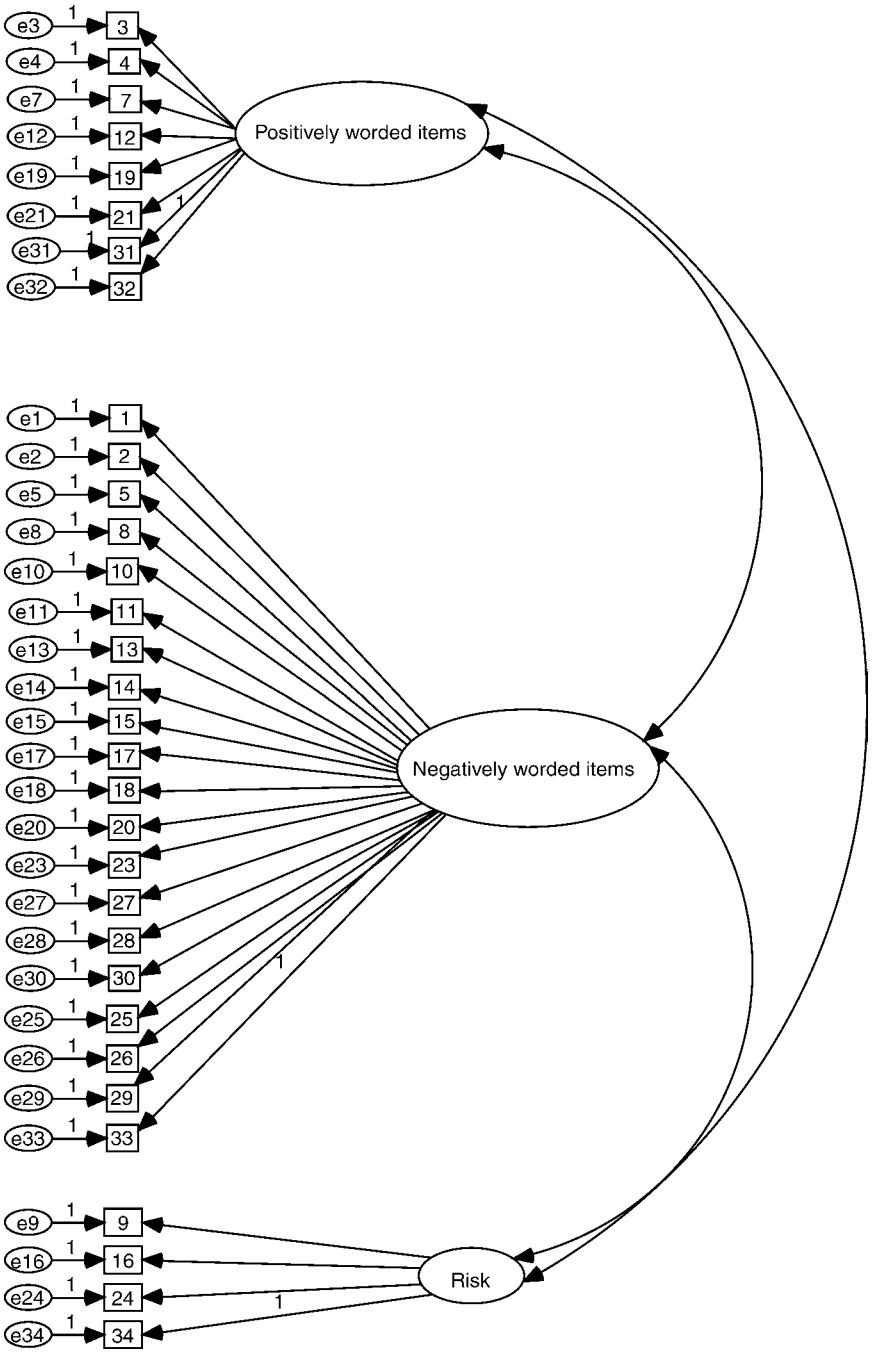


**Figure 1.** (*Continued*) Model 5: positively and negatively worded item groups, and risk correlated first-order latent variables.

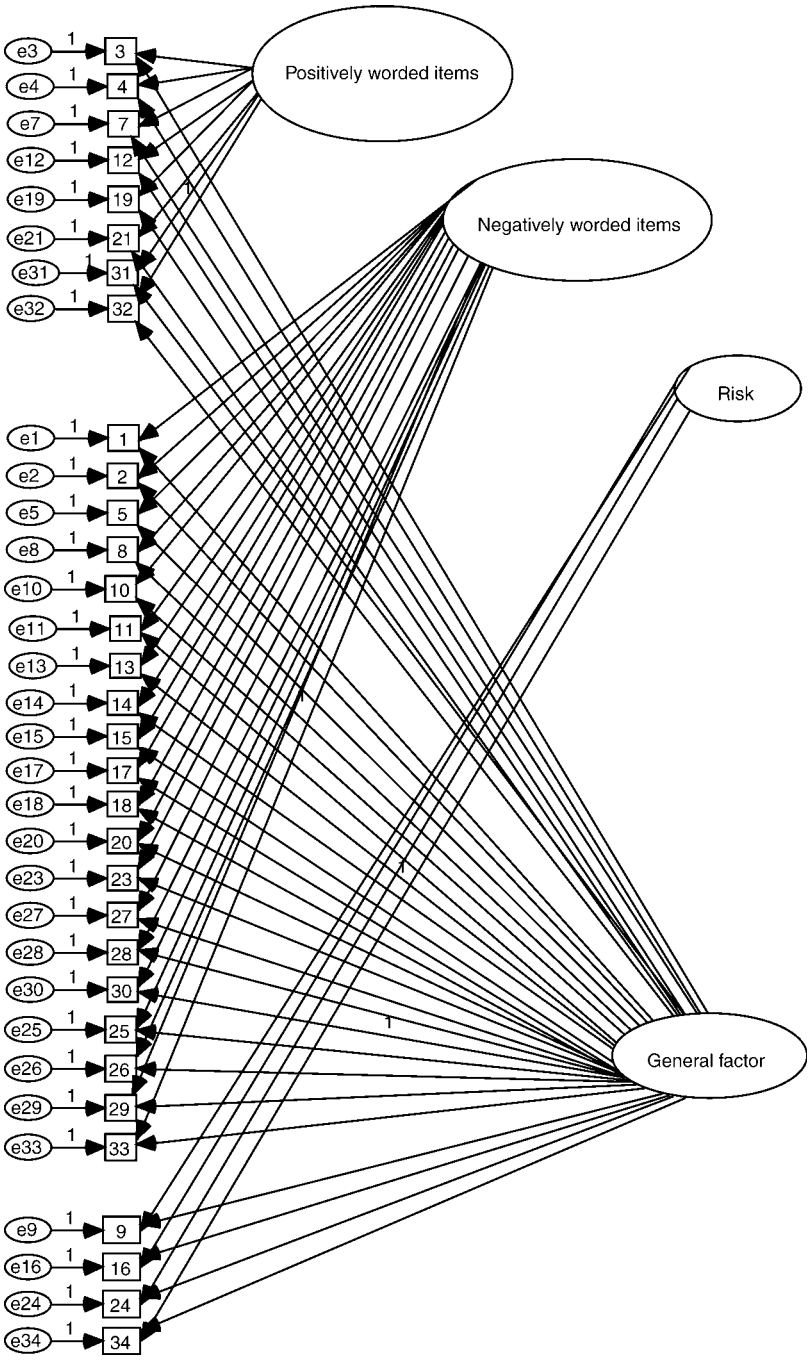


**Figure 1.** (*Continued*) Model 6: a nested factors first-order general factor solution using positively and negatively worded item groups and risk, as first-order residualized latent variables.

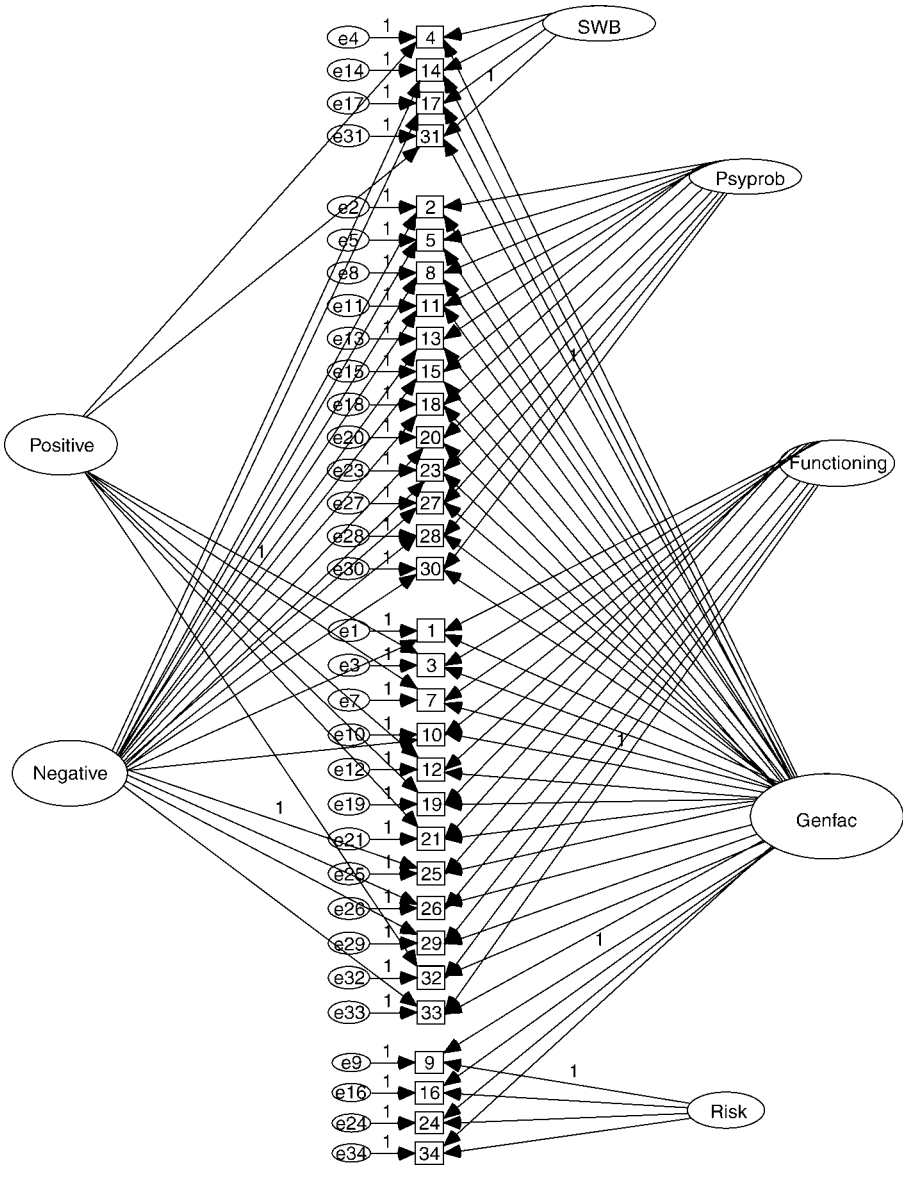


**Figure 1.** (*Continued*) Model 7: a nested factors first-order general factor model with four residualized latents (Model 4b) and with two method latents of positively and negatively worded items.

statistics were still relatively poor. It was decided that for all further modelling, these two risk items would remain excluded.

The correlations for Model 1b were indicative of complete redundancy for the subjective well-being and psychological problems latent variables. This correlation exceeded 1.0 in Amos 5 (1.01) and was constrained to 1.0 in SEPATH. The correlation between subjective well-being and functioning was also very high (.95), with psychological problems and functioning correlating at .91. Risk latent variable

**Table 3.** The fit statistics associated with the ten models examining the structure of the CORE-OM questionnaire item set

| Model # | Chi square | *df* | *p* | Mc | RMSEA | CFI | SRMR |
|---|---|---|---|---|---|---|---|
| 1a | 5,551.34 | 521 | <.0001 | .242 (.227, .257) | .074 (.072, .075) | .839 | .053 |
| 1b | 4,958.64 | 458 | <.0001 | .279 (.263, .295) | .075 (.073, .076) | .852 | .052 |
| 2 | 5152.19 | 461 | <.0001 | .262 (.247, .278) | .076 (.075, .078) | .846 | .052 |
| 3a | 5,278.97 | 463 | <.0001 | .246 (.232, .262) | .078 (.076, .079) | .842 | .053 |
| 3b | 7,469.49 | 464 | <.0001 | .122 (.113, .131) | .095 (.094, .097) | .770 | .063 |
| 4a | 3,808.50 | 435 | <.0001 | .392 (.372, .412) | .066 (.064, .067) | .889 | .045 |
| 4b | 3,651.97 | 432 | <.0001 | .405 (.385, .425) | .065 (.063, .066) | .894 | .042 |
| 5 | 4,517.55 | 461 | <.0001 | .322 (.304, .340) | .070 (.068, .072) | .867 | .050 |
| 6 | 3,270.50 | 432 | <.0001 | .468 (.447, .489) | .059 (.058, .061) | .907 | .038 |
| 7 | 2,458.28 | 404 | <.0001 | .589 (.566, .611) | .051 (.049, .053) | .933 | .035 |

*df* = degrees of freedom, *p* = probability of occurrence, Mc = McDonald non-centrality index, RMSEA = root mean squared error of approximation, CFI = comparative fit index, SRMR = standardized root mean squared residual, PWNW = positively worded and negatively worded latents. Figures in brackets for the Mc and RMSEA close-fit indices are the 90% upper and lower confidence intervals, respectively. Model 1a: complete 34-item CORE-OM model, four latent variables, all four are correlated; Model 1b: 32-item CORE-OM model, four latent variables, all four are correlated, two Risk items 6 and 22 deleted; Model 2: 32-item subjective well-being and psychological problems, represented as a single latent – with functioning and risk as correlated latents; Model 3a: 32-item single general factor latent composed of the 28 non-risk items, with risk as a correlated latent; Model 3b: 32-item single general factor latent composed of all 32 items; Model 4a: 32-item nested factors first-order general factor latent, three residualized first-order latents, and risk as a correlated latent; Model 4b: 32-item nested factors first-order general factor latent, four residualized first-order latents; Model 5: 32-item PW NW model, positively and negatively worded latents, with risk as a correlated first-order latent variable; Model 6: 32-item PW NW model, nested factors first-order general factor latent, using positively and negatively worded method latents, and risk, as residualized first-order latent variables; Model 7: 32-item nested factors first-order general factor latent, four residualized first-order CORE-OM latents, with positively and negatively worded method latents.

correlations with subjective well-being, psychological problems, and functioning were .58, .57, and .61, respectively.

In order to establish that the risk latent variable was indeed coherent as a latent variable, a simple SEM single latent variable model was fitted to the four risk items. This 4-item model produced an exact-fit chi-square of 45.584, $df = 2, p < .0001$. The close-fit indices were: Mc = .99, RMSEA = .10, CFI = .99, and SRMR = .02. The largest observed standardized residual correlation between the model-implied correlation matrix and observed correlation matrix was .05. Given these close-fit results, and the trivial size of the largest standardized residual, this was taken to be reasonable evidence that the risk 4-item model could indeed be considered a latent variable consisting of four manifest items, 9, 16, 24, and 34, with standardized paths (loadings) of .85, .81, .78, and .56, respectively.

Given the identity correlation between subjective well-being and psychological problems in Model 1b, a new model (Model 2) was tested which combined the items from these latent variables. Three latent variables were fitted to the data in Model 2, the original variable of functioning, the new variable called psychological problems, and the 4-item risk variable. The model fit was closely similar to Model 1b (Table 3). Given the high correlation of .96 between the functioning and psychological problem

latent variable in Model 2, a general factor model was created, Model 3a, where all items originally modelled as being manifest indicators of three latent variables (subjective well-being, psychological problems, and functioning) were now considered as manifest indicators of a single general factor latent. The risk latent was modelled as a distinct but correlated second latent variable. The fit statistics for this model were similar to those for the previous models (Table 3). In case the reason for the poor fit was due to the risk latent being modelled as a separate variable, Model 3b was created in which the reduced pool of 32 CORE-OM items were considered manifest indicators of a single general-factor latent. Table 3 shows that this was the worst-fitting model tested so far.

*Test of a nested factors first-order and general factor CORE-OM model*

In order to test for the possibility that the CORE-OM structure might be better modelled by the addition of a general factor latent, Model 4a was created. This utilized a nested factors model for CORE-OM, where the three latent variables of subjective well-being, psychological problems, and functioning were constructed as 'residualized' latent variables, with a general-factor latent modelled as a separate global latent variable, and the risk variable modelled as a separate latent which was correlated with the general factor latent. Examining the fit of Model 4a to the data resulted in an improvement in close-fit indices compared with Model 3a. Model 4b took the nested factors model one step further by incorporating the risk variable as a residualized first-order latent, with the general-factor latent variable modelled as causal for covariance amongst the reduced pool of 32 CORE-OM items. This model was marginally better than that for Model 4a (Table 3).

*Tests of a method factors model for positively and negatively worded items*

Having completed modelling of the published CORE-OM score key, the hypothesis that the structure of CORE-OM might be more closely modelled by positively and negatively worded item method factors was tested.

Model 5 consisted of the two latent variables for positively worded items (PW) and negatively worded items (NW) together with the latent of risk. Given the high correlations among the four content domains, it seemed likely that the PW and NW latents would also be correlated, thus the correlation between them was estimated in this model. The risk latent was also allowed to correlate with these latents. This model gave better fit statistics than Models 1a–3b, but worse than Models 4a and 4b (Table 3).

The latent variable correlations between PW and NW, PW and risk, and NW and risk were .84, .54, and .59, respectively. This indicated that it might be worth fitting a nested factors model to these data, with residualized latent variables of PW, NW, and risk, and a global general factor latent with a path to each of the 32 manifest variables. Model 6 in Fig. 1 shows the model schematic. This model had better fit statistics than any of the previous models (Table 3).

*A multi-method multi-trait nested factors model*

That there may be method variance confounding the assessment of the four CORE-OM first-order domain latents was examined in Model 7. This comprised Model 4b (which included a general factor latent) with the addition of two method latent variables

for PW and NW items. This model was the best fitting of all 10 models, with statistically significant improvements in fit compared with Models 4b and 6. The goodness-of-fit criteria were met for RMSEA and SRMR, and this model came closest of all of the models to meeting the goodness-of-fit criteria for the other close-fit statistics. Table 4 presents the standardized paths from Model 7 for the 32 CORE-OM items.

### *Model comparisons*

The most important model comparisons are shown in Table 5. All of the chi-squared tests resulted in differences that were highly significant, $p < .000001$.

**Table 4.** Model 7 standardized path values: two orthogonal method factors, four first-order orthogonal factors, and one first-order nested general factor

| | Method factors | | CORE original factors | | | | |
|---|---|---|---|---|---|---|---|
| Item | Positive | Negative | SWB | PsycProb | Function | Risk | General factor |
| 1 | | −.08 | | | .07 | | .73 |
| 2 | | −.25 | | .15 | | | .66 |
| 3 | .19 | | | | .20 | | .35 |
| 4 | .28 | | .16 | | | | .67 |
| 5 | | −.17 | | −.01 | | | .59 |
| 7 | .32 | | | | −.01 | | .61 |
| 8 | | −.17 | | .17 | | | .26 |
| 9 | | | | | | .69 | .48 |
| 10 | | −.10 | | | .14 | | .55 |
| 11 | | −.40 | | .20 | | | .63 |
| 12 | .42 | | | | .03 | | .62 |
| 13 | | .12 | | .31 | | | .57 |
| 14 | | .11 | .02 | | | | .67 |
| 15 | | −.28 | | .30 | | | .58 |
| 16 | | | | | | .73 | .39 |
| 17 | | −.09 | −.13 | | | | .80 |
| 18 | | −.04 | | .16 | | | .46 |
| 19 | .29 | | | | .13 | | .27 |
| 20 | | −.03 | | .01 | | | .69 |
| 21 | .36 | | | | .05 | | .52 |
| 23 | | −.00 | | −.04 | | | .82 |
| 24 | | | | | | .56 | .57 |
| 25 | | .07 | | | .59 | | .49 |
| 26 | | .11 | | | .41 | | .51 |
| 27 | | .13 | | −.09 | | | .82 |
| 28 | | .37 | | .59 | | | .56 |
| 29 | | .02 | | | .24 | | .50 |
| 30 | | .09 | | −.06 | | | .52 |
| 31 | .35 | | .10 | | | | .47 |
| 32 | .46 | | | | .05 | | .49 |
| 33 | | .09 | | | .53 | | .48 |
| 34 | | | | | | .50 | .29 |

SWB = subjective well-being; PsycProb = psychological problems; Function = functioning.

**Table 5.** Model comparisons

| Models compared (the first of each pair has significantly better fit) | Chi square | Degrees of freedom |
|---|---|---|
| CORE-OM (1b) vs. SGF + Risk(3a) | 320.33 | 5 |
| CORE-OM + GF(4b) vs. CORE-OM(1b) | 1,306.67 | 26 |
| MF + Risk(5) vs. CORE-OM(1b) | 441.09 | 3 |
| CORE-OM + GF(4b) vs. SGF + Risk (3a) | 1,627.00 | 31 |
| CORE-OM + GF(4b) vs. MF + Risk(5) | 865.58 | 29 |
| MF + GF +Risk (6) vs. MF + Risk(5) | 1,247.05 | 29 |
| MMMT + GF(7) vs. MF + GF + Risk(6) | 812.22 | 28 |
| MMMT + GF(7) vs. CORE-OM + GF(4b) | 1,193.69 | 28 |

SGF = single general factor latent; GF = general factor latent; MF = method factor latents; MMMT + GF = multi-method, multi-trait model with general factor latent.

## Discussion

This cross-sectional study has compared a range of structural models for CORE-OM based on data from a large sample of clients in psychotherapy and counselling services. In addition, the psychometric quality of two CORE-OM scoring keys has been assessed. The results of the present study are not inconsistent with a picture that has been emerging in the validation work to date, however it is now possible to present a more definitive assessment of the psychometric structure of CORE-OM.

### *Risk domain*

It has been confirmed that when the six risk items are separated into the face valid domains for risk-to-self and risk-to-others, the four risk-to-self items form a single latent variable, leaving the risk-to-others items as clinical flags. Furthermore, it has been shown that combining the risk-to-self factor into a general factor for psychological distress, together with all of the other non-risk CORE-OM items (Model 3b), results in a worse fitting model than if the risk-to-self factor is modelled separately (Model 3a).

### *CORE-OM structure*

Representing CORE-OM as a measure of the correlated domains of subjective well-being, psychological problems, and functioning, together with a correlated risk-to-self latent (excluding items 6 and 22) resulted in a poor fit (Model 1b). A model in which the three domains were collapsed into a general factor latent, with risk-to-self as a correlated latent, mirroring the alternative scoring key for CORE-OM, (Model 3a) also resulted in poor fit.

A nested factors model with a general factor latent for the non-risk items and four residualized first-order latents for risk-to-self, and the other three domains (Model 4b) resulted in a noticeably better fit than Models 1b or 3a, indicating that the four CORE-OM domain factors and the general factor both make a contribution to model fit beyond the contribution of either the general factor or the domain factors alone.

Previous validation work using exploratory factor analysis failed to find a factor structure that replicated the CORE-OM domain score key, but found separate factors for positively and negatively worded items (Evans *et al.*, 2002), suggesting that method factors might have an important part to play in explaining the covariance amongst CORE-OM items. Consistent with this finding, Model 5 comprising three correlated

first-order latents for risk-to-self, and the PW and NW method factors, resulted in better fit statistics than the first-order CORE-OM domains model (i.e. Model 1b).

Furthermore, Model 6, a nested factors model with a general factor latent, and residualized latents for the two method factors and risk-to-self resulted in a better fit than Model 4b, comprising a general factor latent and the four residualized latents for the CORE-OM domains.

Whilst the method factors appear to account for more variance than the CORE-OM domains, the above findings suggest that method factors, CORE-OM domains, and a general factor for psychological distress all make a contribution in accounting for covariation between CORE-OM items. This was tested in a multi-method, multi-trait, nested factors model, which was found to be the best fitting model of all (Model 7). This result suggests that the effect of the methods factors identified in CORE-OM needs to be controlled in any future test development by balancing the direction of items across subscales.

Whilst the CORE-OM domains account independently for some of the variance in Model 7, there is relatively little differentiation between the three non-risk CORE-OM domains as evidenced by the high correlations between them, their strength of association with the general factor for psychological distress, the relatively greater influence of methods factors in SEM, and the poor scale quality results when CORE-OM is scored for the four domains in these cross-sectional data. As mentioned in the Introduction, as yet there is no empirical evidence that they respond differentially over time in therapy, as would have been predicted by the phase model of psychotherapy change (Shapiro *et al.*, 2001).

Inspection of the CORE-OM items suggests that the high covariance of non-risk domains with each other and with the general factor for psychological distress might be due in part to item similarity. A high proportion of the items for the three non-risk scales include similarly worded feelings and emotions phrases. For example, for the functioning domain the following phrases appear, 'felt too much for me', 'felt able to cope', 'been happy', 'felt warmth', 'felt criticized', 'been irritable' and 'felt humiliated', for the problems domain, 'felt tense, anxious', 'felt totally lacking in energy', 'felt panic or terror', 'felt despairing, hopeless', 'felt unhappy' and 'thought I am to blame', and for the well-being domain, 'felt overwhelmed' and 'felt OK'. However, whilst covariance between the domains might be constrained by reducing item similarity, it is appropriate that these constructs would still have significant variance in common with this general factor, given that CORE-OM is a psychological therapies outcome measure.

Inspection of Table 4 shows that the four harm-to-self risk items have substantial loadings on the risk factor in addition to moderate loadings on the general factor, supporting the decision to treat these items as a scale. By contrast, the subjective well-being items do not have meaningful loadings on the subjective well-being factor, although the two positively worded subjective well-being items load on the positive method factor.

Only the three items 13, 15 and 28 have meaningful loadings on the psychological problems factor; these items are associated with panic, terror and unwanted intrusive thoughts and images, that is, powerful subjective psychological disturbance. The three items, 25, 26 and 33 also have meaningful loadings on the functioning factor; these items are to do with relationships with other people, such as feeling humiliated or criticized by others and having no friends.

### *Psychological symptoms and functioning*

Of particular interest is whether the covariance between measures of psychological symptoms and functioning could in principle be reduced, given that the specific

measurement of reduction in symptoms and improvement in functioning had been of importance to therapists and managers when CORE-OM was developed.

Other authors have attempted to operationalize these variables. Mundt, Marks, Shear, and Greist (2002) have shown that the self-report Work and social adjustment scale, which is specifically designed to measure functional impairment in work, home management, social leisure, private leisure, and relationships, discriminates patients suffering from depression and obsessive-compulsive disorder according to symptom severity. However, correlations with standard symptom measures were of the order of .75, which is of the same order as that between the CORE-OM non-risk domains.

The Health of the Nation Outcomes Scale (HoNOS) is used widely in the UK as an outcomes measure in mental health services (Wing, Beevor, & Curtis, 1998). Only three of the 12 HoNOS scales have been shown to discriminate for outcomes in psychological therapy out-patient populations (Audin, Margison, Mellor-Clark, & Barkham, 2001). These were depressed mood, mental and behavioural problems, and problems with relationships. There was a floor effect for scores on the other HoNOS scales, some of which are related to functioning, including activities of daily living, living conditions, occupation and activities, and behavioural problems. It seems that the level of dysfunction in the psychotherapy and counselling population is too low to be detected by many of the HoNOS scales, which have been shown to have greater utility for psychiatric populations with severe and enduring mental illness (Orrell, Yard, Handysides, & Schapira, 1999).

However, in a review, Mintz, Mintz, Arruda, and Hwang, (1992) demonstrated differential rates of recovery for occupational functioning and psychological distress in depressed patients. Reduction in self-reported psychological distress on the BDI (Beck *et al.*, 1988) occurred within a few weeks of treatment with antidepressants, whereas improvement in work-related functioning took longer. The unique feature of this research was that measures of functioning, although self-reported, were behavioural, including sickness absence, involvement in conflicts at work and work performance.

These studies, taken together with the results of the analyses of the CORE-OM, suggest three reasons for the difficulties in operationalizing a distinction between problems (symptoms) and functioning in psychotherapy and counselling out-patient populations.

Firstly, a significant part of the clinical evidence for psychological distress is to do with aspects of functioning, such as lack of energy and social avoidance, hence these variables might be expected to be at least moderately correlated.

Secondly, the degree of functional impairment in psychotherapy and counselling populations is relatively low compared with that in populations of patients with severe and enduring mental illnesses for whom functioning has been reliably measured using other methods, and therefore might need sensitive measures in order to be detectable.

Thirdly, questionnaires such as HoNOs that do successfully operationalize a distinction between functioning and psychological symptoms include indicators that are not self-reported and/or go beyond the realm of symptoms to include objective functional problems. Whilst there is a floor effect for HoNOS scales in psychotherapy out-patient populations, because these scales are designed to detect relatively severe dysfunction, the review by Mintz *et al.* (1992) suggests that a way forward may be to write items for functioning that carefully reflect actual behaviours and coping strategies in daily life, without reference in the wording to the distress/well-being that might accompany them, thereby minimizing covariance between measures of functioning and psychological symptoms. The item loadings in Table 4 (discussed above) confirm that

specific measurement of perception of the quality of relationships and social support may be a particularly useful dimension in measures of psychological functioning.

These considerations with respect to the CORE-OM domains are of importance for future research and scale development, but the utility of CORE-OM has already been demonstrated as a widely used benchmarking measure and reliable indicator of change in psychotherapy research and practice. The scoring method that has proved most useful in this regard is that in which all 28 non-risk items are scored as one scale and the risk items as the other. This research confirms that the scale quality of CORE-OM when scored in this way is satisfactory.

## Acknowledgements

We thank Janice Connell for compiling the original data set. Michael Barkham was supported by the Research and Development Levy from Leeds Community and Mental Health Trust.

## References

Arbuckle, J. L. (2003). *Amos Structural Equation Modeling system, version 5.0*. Chicago: SPSS.

Audin, J., Margison, F., Mellor-Clark, J., & Barkham, M. (2001). Value of HoNOS in assessing patient change in NHS psychotherapy and psychological treatment services. *British Journal of Psychiatry*, *178*, 561–566.

Barkham, M., Evans, C., Margison, F., McGrath, G., Mellor-Clark, J., Milne, D., & Connell, J. (1998). The rationale for developing and implementing core batteries in service settings and psychotherapy outcome research. *Journal of Mental Health*, *7*, 35–47.

Barkham, M., Hardy, G. E., & Startup, M. (1996). The IIP-32: Development of a short version of the Inventory of Interpersonal Problems. *British Journal of Clinical Psychology*, *35*, 21–35.

Barkham, M., Margison, F., Leach, C., Lucock, M., Mellor-Clark, J., Evans, C., Benson, L., Connell, J., Audin, K., & McGrath, G. (2001). Service profiling and outcomes benchmarking using the CORE-OM: Towards practice-based evidence in the psychological therapies. *Journal of Consulting and Clinical Psychology*, *69*, 184–196.

Barrett, P., Kline, P., Paltiel, L., & Eysenk, H. J. (1996). An evaluation of the psychometric properties of the Concept 5.2 Occupational Personality Questionnaire. *British Journal of Occupational and Organizational Psychology*, *69*, 1–19.

Beck, A. T., Epstein, N., Brown, G., & Steer, R. A. (1988). An inventory for measuring clinical anxiety: Psychometric properties. *Journal of Consulting and Clinical Psychology*, *56*, 893–897.

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Manual for the Beck Depression Inventory – second edition (BDI-II)*. San Antonio, TX: Psychological Corporation.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry*, *4*, 561–571.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *2*, 238–246.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334.

Derogatis, L. R. (1983). *SCL-90-R: Administration, scoring and procedures: Manual II*. Towson, MD: Clinical Psychometric Research.

Derogatis, L. R., & Melisaratos, N. (1983). The Brief Symptom Inventory: An introductory report. *Psychological Medicine*, *13*, 595–605.

Evans, C. E., Mellor-Clark, J., Margison, F., Barkham, M., Audin, K., Connell, J., & McGrath, G. (2000). CORE: Clinical outcomer in routine evaluation. *Journal of Mental Health*, *9*, 247–255.

Evans, C., Connell, J., Barkham, M., Margison, F., McGrath, G., Mellor-Clark, J., & Audin, J. (2002). Towards a standardised brief outcome measure: Psychometric properties and utility of the CORE-OM. *British Journal of Psychiatry*, *180*, 51–60.

Feldt, L. S. D., Woodruff, D. J., & Salih, F. A. (1987). Statistical inference for coefficient alpha. *Applied Psychological Measurement*, *11*(1), 93–103.

Goldberg, D. P., & Hillier, V. G. (1979). A scaled version of the General Health Questionnaire. *Psychological Medicine*, *9*, 139–145.

Gustafsson, J., & Balke, G. (1993). General and specific abilities as predictors of school achievement. *Multivariate Behavioral Research*, *28*, 407–434.

Horowitz, L. M., Rosenberg, S. E., Baer, B. A., Ureno, G., & Villasenor, V. S. (1988). Inventory of personal problems: Psychometric properties and clinical applications. *Journal of Consulting and Clinical Psychology*, *56*, 885–892.

Howard, K. I., Lueger, R. J., Maling, M., & Martinovich, Z. (1993). A phase model of psychotherapy: Causal mediation. *Journal of Consulting and Clinical Psychology*, *61*, 678–685.

Hu, L., & Bentler, P. M. (1999). Cut off criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55.

Joyce, A. S., Ogrodniczuk, J., Piper, W. E., & McCallum, M. (2002). A test of the phase model of psychotherapy change. *Canadian Journal of Psychiatry*, *47*, 759–766.

McDonald, R. P. (1989). An index of goodness-of-fit based on non-centrality. *Journal of Classification*, *6*, 97–103.

Mellor-Clark, J., & Barkham, M. (2000). Quality evaluation: Methods, measures and meaning. In C. Feltham & I. Horton (Eds.), *Handbook of counselling and psychotherapy* (pp. 255–270). London: Sage.

Mellor-Clark, J., Barkham, M., Connell, J., & Evans, C. (1999). Practice-based evidence and the need for a standardised evaluation system: Informing the design of the CORE system. *European Journal of Psychotherapy Counselling and Health*, *3*, 357–374.

Mintz, J., Mintz, L. I., Arruda, M. J., & Hwang, S. S. (1992). Treatments of depression and the functional capacity to work. *Archives of General Psychiatry*, *49*, 761–768.

Mundt, J. C., Marks, I. M., Shear, M. K., & Greist, J. H. (2002). The Work and Social Adjustment Scale: A simple measure of impairment in functioning. *British Journal of Psychiatry*, *180*, 461–464.

Orrell, M., Yard, P., Handysides, J., & Schapira, R. (1999). Validity and reliability of the Health of the Nation Outcome Scales in psychiatric patients in the community. *British Journal of Psychiatry*, *174*, 409–412.

Shapiro, D. A., Leach, C., Diggle, P., Lucock, M., Iveson, S., & Barkham, M. (2001, March). *Testing the phase model with the CORE outcome measure in routine psychological therapies practice*. Paper presented at the UK & Continental European Joint Meeting of the Society for Psychotherapy Research, Leiden, The Netherlands.

StatSoft, Inc. (2003). *STATISTICA (data analysis software system), version 6*. Tulsa, OK: Author.

Steiger, J. H. (1980a). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *2*, 245–251.

Steiger, J. H. (1980b). Testing pattern hypotheses on correlation matrices: Alternative statistics and some empirical results. *Multivariate Behavioural Research*, *15*, 335–352.

Steiger, J. H. & Lind, J. C. (1980, May) *Statistically-based tests for the number of common factors*. Paper presented at the Annual Spring Meeting of the Psychometric Society in Iowa City, IA.

Strupp, H. H., Horowitz, L. M., & Lambert, M. J. E. (1997). *Measuring patient changes in mood, anxiety, and personality disorders: Toward a core battery*. Washington DC: APA.

Waskow, I. E. (1975). Selection of a core battery. In M. B. Parloff (Ed.), *Psychotherapy change measures*. Washington, DC: U.S. Government Printing Office.

Wing, J. K., Beevor, A. S., & Curtis, R. H. (1998). Health of the Nation Outcome Scales (HoNOS). Glossary for HoNOS score sheet. *British Journal of Psychiatry*, *174*, 432–434.

*Received 27 March 2003; revised version received 17 January 2005*