Clinical Psychology and Psychotherapy
Clin. Psychol. Psychother. **17**, 51–62 (2010)
Published online 2 September 2009 in Wiley InterScience (www.interscience.wiley.com). **DOI**: 10.1002/cpp.649

Assessment

Mokken Scaling and Principal Components Analyses of the CORE-OM in a Large Clinical Sample

A. Bedford,[1,2]* R. Watson,[3] J. Lyne,[2] J. Tibbles,[4] F. Davies[1] and I. J. Deary[5]

[1] *Department of Psychological Therapies, The Old Chapel, Bootham Park, York, UK*
[2] *Department of Psychology, The University of York, Heslington, York, UK*
[3] *School of Nursing and Midwifery, The University of Sheffield, Samuel Fox House, Northern General Hospital, Sheffield, UK*
[4] *NCYPE, Lingfield, Surrey, UK*
[5] *Medical Research Council Centre for Cognitive Ageing and Cognitive Epidemiology, Department of Psychology, The University of Edinburgh, Edinburgh, UK*

In a sample of 543 adult National Health Service (NHS) patients referred to a Psychological Therapies Service, the responses to the Clinical Outcomes in Routine Evaluation-Outcome Measure (CORE-OM) self-report questionnaire were examined using conventional principal components analysis (PCA) and a unique application of Mokken Scaling Procedure (MSP). Following the theoretical views of G. A. Foulds, it was suggested that some items more properly belong to the universe of attitudes and traits rather than that of symptoms and states. Accordingly, the analyses were carried out both with and without the CORE-OM Risk domain items. Both PCAs produced a very large first component of Psychological distress, while the small second component differs. With all items included, the second component was of Risk. With the risk items excluded, the second component was now Functioning. The MSP results, respectively, were of a five-item scale of Functioning (impaired by depression) and on the second analysis, a five-item Functioning scale (impaired by anxiety). There was discussion on the criteria for item selection, the time scale specified in questionnaire instructions and the optimum number of items required for a symptom scale. It was concluded that the CORE-OM item pool did not conform to its purported face validity domains and subdomains, but predominantly constitutes a large Psychological distress scale with considerable item redundancy. Copyright © 2009 John Wiley & Sons, Ltd.

*Correspondence to: A. Bedford, Department of Psychological Therapies, The Old Chapel, Bootham Park, York, YO30 7BY UK.
E-mail: alanbedford65@hotmail.com

Keywords: CORE-OM, Mokken Scaling Procedure (MSP), Principal Components Analysis (PCA)

INTRODUCTION

An epidemiological survey by the Office of National Statistics of the general population conducted in Great Britain suggested that 1 in 6 adults had a neurotic disorder; i.e., such specific conditions as anxiety, depression and phobias (Singleton, Bumpstead, O'Brien, Lee, & Meltzer, 2000). With such a high prevalence rate, there is an obvious need for valid measures of general psychological distress for diagnosis and treatment evaluation in clinical settings. The intention of the Clinical Outcomes in Routine Evaluation-Outcome Measure (CORE-OM) creators was, following the sought opinions of clinicians, to establish a questionnaire that could become the United Kingdom's (UK) 'Gold Standard', with relevant local standardization. This could replace the plethora of other, often imported, psychometric tests. The CORE-OM has since become an extremely widely used self-report questionnaire in the UK National Health Service (NHS).

The aim of the present paper is to evaluate the structure of the CORE-OM using conventional principal components analysis (PCA) and additionally through the first application of the Mokken Scaling Procedure (MSP) to a CORE-OM database. In so doing, it complements an earlier study using Structural Equation Modelling (Lyne, Barrett, Evans, & Barkham, 2006), thereby filling a gap in the literature on a widely used UK questionnaire, for which there is only scant published knowledge of its structure.

The CORE-OM (Evans et al., 2000; Evans et al., 2002) was included as one of a range of outcome measures in a routine service evaluation at an NHS Primary Care Trust Psychological Therapies Department for working-age adults. The CORE-OM consists of 34 statements to be answered 'over the last week', with five choices ranging from 'not at all' to 'most or all of the time'. These individual items, according to the test authors, summate into four domains and nine subdomains as shown in Table 1.

It will be noted that there is a gross imbalance between the numbers of items in the domains and subdomains, compared to most, if not all, psychological inventories. This raises the issue as to the relative value of such groupings. With regard to an analysis aimed at discovering the factorial structure of the CORE-OM, Evans et al., (2002) stated that 'a first component accounts for a large proportion of the variance, but a three-component structure that separates problems, risk items and positively scored items may be worthy of further exploration . . .'. (Presumably 'positively *worded* items' was intended).

In a study consisting of 2140 adult NHS patients receiving psychological therapy in the UK, Lyne, et al. (2006) compared a range of psychometric models for the questionnaire using Structural Equation Modelling. They found poor fit for the CORE-OM four domains model. Scale Quality (a measure of signal to noise ratio between scales) was unacceptably low for this model. An alternative model suggested by a factor analysis under-

Table 1. CORE-OM structure of domains and sub-domains of items as devised by the test authors

Number of items	Domains	Subdomains and number of items
4	Subjective well-being	Not applicable
12	Problems or symptoms	Anxiety (4) Depression (4) Physical (2) Trauma (2)
12	Functioning	General (4) Close relationships (4) Social relationships (4)
6	Risk	Harm to self (4) Harm to others (2)

DOI: 10.1002/cpp

taken by Evans et al. (2002) gave three factors: these were risk, positively worded items and negatively worded items, but this model also resulted in a poor fit to the data. The best fit was achieved with a complicated multimethod (positive and negative wording of items), multitrait (the other three domains together with a single psychological distress factor) model that was unscorable in practice. The authors concluded that CORE-OM might best be scored as a single 28-item psychological distress scale, with a separate four-item risk-to-self scale. In other words, Lyne et al. (2006) found no empirical evidence to support the purported domains and subdomains of the test authors.

On a broader canvas, Foulds (1965) suggested that in general it might prove useful to distinguish between the symptoms and signs of what he termed personal illness, and of personality traits and attitudes, as these belonged to logically different universes of discourse. His differentiae are that traits and attitudes are egosyntonic, relatively stable and scale scores are normally distributed in the general population. By contrast, symptoms and states are distressing, relatively transient and have markedly positively skewed distributions in the general population. It might also be expected that these aspects of the person would show differential change over time and after undergoing therapy. It is improper and meaningless, therefore, to summate scores of such attributes from these different logical universes of discourse (Foulds, 1965, 1971, 1976). When the item content of the CORE-OM is examined, a similar case can be made for a separation between the majority of items in the 'Problems or Symptoms' and 'Subjective Well-Being' (negated) domains, as representing symptoms-states, and the rest of the items. In particular, the 'Risk' domain items, concerning 'Harm to self' and 'Harm to others' are intended to tap what in Foulds' terms would be intropunitive (inwardly directed hostility) and extrapunitive (outwardly directed hostility) attitudes or traits (Caine, Foulds, & Hope, 1967). The 'Functioning' domain, which focuses predominantly on personal relationships, may occupy an intermediate conceptual position between the symptom and the 'Risk' items. However, the minimal exclusion of at least the 'Risk' domain items could lead to a better understanding of the CORE-OM questionnaire symptom structure.

The intention of this report is to assess the factorial structure of the CORE-OM using conventional PCA and by additionally employing the first application of the MSP to a CORE-OM database.

The MSP will be used to discover whether there is a hierarchical scale within the CORE-OM's items. Because the MSP is a much less familiar analytic technique than the various forms of factor analysis, it requires additional introductory explanation.

Hierarchical scales can best be explained by example; for instance, the case of a football league table. A football league table is a measure of the latent trait 'ability at football' and the position of a team in the league table represents its ability at football and, thereby, the likelihood that it has beaten almost all the teams below it in the league and been beaten by almost all of those above it in the league. Certainly, at the end of the football season, when the league position is finally fixed, in terms of overall league performance—the proxy measure of the latent trait 'ability at football'—a team with an intermediate position in the league will have beaten almost all the teams below it and been beaten by almost all the teams above it; the team at the top of the league will have beaten almost all the teams below it and vice versa for the bottom of the league. Clearly, the position in the league at any point during the season and even at the end will only be a general indication of the outcome of specific matches as it is perfectly possible, for example, for the team at the bottom of the league to have beaten the team at the top of the league in their matches—but, nevertheless, unlikely. Therefore, the position in the league is not a perfect indicator of performance in any particular game—each game being a stochastically independent event. Nevertheless, position in the league will be an indicator, indeed a measure, of 'ability at football'. We may think of items within questionnaire scales as being like teams within the same league table.

The recent demonstration of a hierarchical scale of psychological distress composed of items from the 30-item General Health Questionnaire (GHQ) (Watson, Deary, & Shipley, 2008a), and other hierarchical scales in commonly used psychological inventories, i.e., the NEO Five Factor Inventory (NEO-FFI) and the Eysenck Personality Inventory (EPI) (Watson, Deary, & Austin, 2007; Watson, Roberts, Gow, & Deary, 2008b), prompted the search for a hierarchy of items in the CORE-OM. Hierarchies in scale items emanate from item-response theory (IRT) and represent a complementary way by comparison with factor analysis and PCA of looking at the relationship between item scores and scale total scores. In this way, an individual's score on a questionnaire can be related

DOI: 10.1002/cpp

to the underlying theoretical construct that is being measured (Hulin, Drasgow, & Parsons, 1983). Scales, thereby, become more interesting from a theoretical perspective because the score of an individual on a latent trait can be related to individual items on an inventory and these actually represent the extent to which the latent trait is present.

The present study, therefore, applies both classical test theory (PCA and internal consistency) and IRT (MSP) to the items of the CORE-OM to investigate its latent dimensions and the possibility of hierarchical sets of items. Such a study could increase both the understanding and utility of the scale.

In the current investigation, Cronbach's coefficient alpha will be employed as an estimate of internal consistency reliability of existing or derived scales of the CORE-OM. Traditionally, an alpha of 0.70 has been regarded as indicating an adequate level of reliability. Cortina (1993) adds that this statistic must be considered in conjunction with the number of items in the particular scale, analogous to the evaluation of correlation size and sample size. When coefficient alpha reaches levels of 0.90 and greater, there is a danger of scales being composed of items that are too similar to each other in content. This leads to a narrow construct being measured, e.g., if all items in an extraversion scale only referred to behaviour at parties. A lot of items then, being highly similar, are likely to be redundant. This is an undesirable property for any scale, as redundant items make no contribution and waste valuable time. The testee, faced with an unnecessarily lengthy scale, may become irritated or bored, resulting in random responses or, even more extremely, non-compliance.

METHOD

Data from 661 adults who had completed the CORE-OM were entered into an SPSS version 15.0 database (SPSS Inc., Chicago, IL, USA). Individuals with any missing data (including demographic data) were removed from the database to prepare the data for MSP analysis, leaving 543 participants with complete CORE-OM data for analysis. These were 160 males with mean age 40.1, standard deviation (SD) 12.8, and 383 females with mean age 37.4, SD 11.6. Almost all had been referred by their general practitioners and had completed the CORE-OM questionnaire before their first clinical appointment. Mokken scaling was carried out using the MSP version 5.0 for Windows (Molenaar & Sijtsma, 2000). The procedure for running the MSP and selecting items has already been described by Watson et al. (2007). Briefly, using the SEARCH facility in the MSP programme, all items were entered into the analysis, setting p at <0.05 and gradually increasing Loevinger's coefficient of homogeneity (H) (to be described below) from 0 to 0.55 in incremental steps of 0.05. Initially, all the items form a single scale until H is approximately 0.3. Thereafter, multiple scales begin to form, and the point at which the largest number of reliable scales (Rho > 0.7; to be explained below) is obtained at approximately H = 0.45, is the point at which those scales can be analysed in more detail using the full diagnostic capabilities of the MSP.

Mokken scaling (Mokken & Lewis, 1982) is based upon IRT and, similar to Rasch scaling (van Schur, 2003), is a technique for establishing hierarchical scales. It is less restrictive in its assumptions than Rasch scaling (Meijer, Sijtsma, & Smid, 1990) but, in common with Rasch scaling, assumes local stochastic independence of items and avoids violations of monotone homogeneity and double monotonicity (Meijer et al., 1990). Local stochastic independence is the assumption that items score as they do due to the level of the latent trait that is being measured and not as a result of a score on any other item in the scale. Monotone homogeneity shows that the score on an item increases as the score on the latent trait increases, and double monotony shows that the item-response curves for the items in the scale do not intersect. Unlike classical test theory (factor analysis and Cronbach's alpha), which uses 'top down' theoretical procedures for item selection, Mokken scaling works through a 'bottom up' procedure; items are selected that conform best to the assumptions of the model and other items are then clustered around these (van Schur, 2003). In IRT, items are positioned in a hierarchy on the basis of their 'difficulty', which, in terms of IRT, means the likelihood of the item being endorsed. More 'difficult' items are less likely to be endorsed and, while indicating a greater presence of the latent trait being measured, will have lower mean values than less 'difficult' items.

Mokken scaling analysis is carried out using the MSP (Sijtsma. Debets & Molenaar, 1990). Applications of the MSP have been explained by Watson and co-workers recently (Watson et al., 2007; Watson et al., 2008a; Watson et al., 2008b). The MSP provides diagnostics that demonstrate whether reliable hierarchies of items exist in a multivariate

DOI: 10.1002/cpp

data set and checks that these items have monotone homogeneity and double monotony. Loevinger's coefficient of homogeneity (H) provides a measure of how often items are found relative to one another in a group of individuals responding to a set of items and a Loevinger's coefficient ≥0.3 is considered to indicate the presence of a hierarchical scale. The *Crit* diagnostic is a combined statistic generated by the MSP, which indicates the extent to which H falls below an acceptable level, how many violations of the Mokken model there are, the size of these violations (van Schur, 2003) and detects violations of monotone homogeneity and double monotonicity. *Crit* ≥80 shows that items should be discarded, while by contrast values ≤40 are ideal. Scale reliability (Rho) is calculated using a method analogous to Cronbach's alpha (Watson et al., 2007). Rho > 0.7 indicates a reliable scale. The probability of obtaining the scale, correcting for the multiple iterations of the MSP programme into account, is also calculated.

Following Foulds' general theoretical position, regarding the symptom/state versus trait/attitude distinction, the CORE-OM was analysed twice: once with all items included and then with the 'Risk' items omitted. On both occasions, first a PCA was carried out, followed by the MSP. Note that in the full text of the CORE-OM, 21 of the 34 items are prefixed by 'I have felt . . .' or 'I have been . . .'.

RESULTS

With all 34 Items Included

The first approach in assessing the structure of the CORE-OM item pool was to adopt the viewpoint and assumptions of the test authors. Accordingly, we allotted the 34 individual items to the domains of 'Subjective Well-Being' (4), 'Problems or Symptoms' (12), 'Functioning' (12) or 'Risk'(6), as explained in the CORE System User Manual, and computed the Cronbach's alpha coefficients. The results by domain were: 'Subjective Well-Being', 0.75; 'Problems or Symptoms', 0.88; and 'Functioning', 0.87. Finally, 'Risk' had a Cronbach's alpha of 0.75, which increased to 0.80 when items 6 ('physically violent to others') and 22 ('threatened or intimidated another person') were excluded. These two items constitute the entire 'harm to others' subdomain. All these values indicate good internal consistency.

While it might seem appropriate at this stage to present descriptive statistics for the domains, this would presume that the structure intended by the test authors had already been confirmed. All that has been established so far is the high (and in two instances, debatably over high) internal consistency of each individual domain, but not their intercorrelations, or the intercorrelations of their constituent items across domains.

Principal Components Analyses

A PCA was carried out with oblimin rotation seeking an oblique solution following the suggestion of Evans et al. (2002). The problem of attempting to interpret a matrix of factor loadings of questionnaire items can be contentious. When confronted with the matrix of the Mood and Anxiety Symptom Questionnaire's 90 items (MASQ; Clark & Watson, 1991), intended as a measure of the controversial tripartite model of adverse mood states, Bedford (1997) had included items only if they satisfied the criteria of: '(a) have a loading of 0.30 or greater . . . and (b) the major loading should be more than 0.20 greater than any cross-loading' p. 125. Keogh and Reidy (2000) used these rules equally successfully to simplify another matrix of MASQ component loadings, and Boschen and Oei (2006) noted these criteria in their multimodel confirmatory factor analyses of the MASQ.

Requested Four-Component Solution. A four-component solution in line with the test authors' assumptions of four distinct domains was requested. This resulted in, apart from the first component having 14 items, the remaining three components having four, two and two items, respectively. That being clearly unsatisfactory, a three-component solution was attempted, in line with the previously mentioned suggestion of Evans et al. (2002).

Requested Three-Component Solution. Component 1 consisted mainly of items, to use the test authors' terminology of the 'Problems or Symptoms' domain, being from their subdomains of 'anxiety' (4), 'trauma' (2), 'depression' (2) and 'physical' (1). Together with the 'Subjective Well-Being' (2) domain and the 'functioning–social relationships' subdomain (2), this component clearly constituted a 13-item Psychological Distress domain, with loadings ranging from 0.47 to 0.79. The second component had three items, all from the 'Risk to self' subdomain, with loadings ranging from 0.74 to 0.81. The third component also only had three loadings. Two items were from the 'Functioning—close relationships' subdomain and

DOI: 10.1002/cpp

one from the 'Subjective Well-Being' domain. These three-item components are not viable as scales because of low reliability and validity and do not appear to have psychological value. Indeed, Kline (1993, p. 38) states that 'it is clear that a reliable test can be made from as few as 10 homogeneous items but this is probably a minimum figure for a good test'. Homogeneity of items in turn implies a clear, distinct unifactorial scale or component. The next logical step was to seek a two-component solution.

Requested Two-Component Solution. In the two-component solution, the first rotated component, with a Cronbach's alpha coefficient of 0.94 and loadings of 0.41–0.80, was composed of 24 items, being 10 of the 'Problems or Symptoms' domain, 10 from the 'Functioning' domain and 4 of the 'Subjective Well-Being' domain (see Table 2). That this is a Psychological distress component can be seen from the highest loadings, which were, for item 27, 'I have felt unhappy' (0.80); item 12, 'I have been happy with the things I have done' (0.77); item 4, 'I have felt O.K. about myself' (0.77); and item 32, 'I have achieved the things I wanted to' (0.77). The latter three statements, of course, were negated. The second component had a Cronbach's alpha of 0.69 (which cannot be improved) and loadings of 0.55–0.71, and was composed entirely of five of the six 'Risk' domain items. (See Table 2.)

The Pearson product–moment correlation between components 1 and 2 was a statistically significant 0.48 ($p < 0.001$). Only the risk component correlated significantly with age, at –0.18 ($p < 0.001$).

Requested One-Component Solution. Finally, for completion, an unrotated one-component solution was sought for all 34 CORE-OM items (see Table 2). Only one item had loading below 0.30, while a further four items were below 0.40, leaving 29 items with loadings ranging from 0.40 to 0.83. This component accounted for 36.0% of the total variance. When the scale was reduced to the 29 items with loadings above 0.40, the component accounted for 40.7% of the total variance.

Mokken Scale Analysis

The scale derived from all 34 CORE-OM items is shown in Table 3, and is composed of six items with acceptable scaling properties (H = 0.43), reliability (Rho = 0.78) and probability ($p = 0.00067$). This scale covers a range of items in terms of difficulty and runs from the relatively mild distress of not having 'achieved the things I wanted to do' (item 32) to having 'made plans to end my life' (item 16).

With only 28 Non-Risk Items

Principal Components Analyses

Requested Three-Component Solution. Components were selected on the conventional basis of examination of the size of eigenvalues and inspection of the scree slope (Cattell, 1966). The first analysis of just 28 CORE-OM items, the six risk items having been excluded, gave a scree slope suggesting a three-component solution (also the test authors' intention), while five components had eigenvalues greater than 1. When the inclusion/exclusion criteria outlined earlier continued to be employed, the third and final component was found to be comprised of only three items (25, 29 and 33), which are three-quarters of the 'Functioning—social relationship' subdomain. Additionally, the latter all contain the words 'other people', which suggested that this component might be a bloated specific (i.e., 'items which are essentially paraphrases of each other' (Kline, 1994). A scale of such items would have high internal consistency, i.e., large Cronbach's alpha, but very low validity.

Requested Two-Component Solution. A two-component solution resulted in 14 and 5 items with Cronbach's alpha coefficients of 0.90 and 0.75, respectively (see Table 2). The first rotated component with loadings ranging from 0.43 to 0.80 consisted of items describing Psychological distress, with the sole exception of item 8 ('aches, pains or other physical problems'). The second component's five-item loadings were from 0.62 to 0.70, four of these being from the 'Functioning' domain and being positively worded. The exception is a negatively worded item 31 ('felt overwhelmed by my problems') from the 'Subjective Well-Being' domain. The Pearson correlation between these two components was a statistically significant 0.58 ($p < 0.001$), while neither correlated significantly with age.

Requested One-Component Solution. For completion, a one-component solution for the 28 items was computed and accounted for 39.8% of the total variance (See Table 2). Apart from item 8 (with a loading of 0.31), item 19 (0.36) and item 3 (0.49), the rest of the items loaded within the 0.50–0.59 band in seven cases, 0.60–0.69 in 11 cases and 0.70–0.79 in four cases. Additionally, three items loaded 0.80

DOI: 10.1002/cpp

Table 2. CORE-OM items abbreviated content, principal components loadings and Mokken scale membership. (n = 543)

Item number	Abbreviated item content	Thirty-four items				Twenty-eight items			
		Two-component solution		One-component solution	Mokken scale items	Two-component solution		One-component solution	Mokken Scale items
		I	II	I		I	II	I	
1	Terribly alone and isolated	71	10	75				75	✓
2	Tense, anxious or nervous	63	13	69		75	1	70	✓
3	Someone to turn to for support +	53	–5	49		8	70	49	
4	OK about myself +	77	–8	71				72	
5	Totally lacking in energy and enthusiasm	68	–1	61				63	
6	Physically violent to others	11	60	19					
7	Able to cope when things go wrong +	64	0	61	✓			62	✓
8	Troubles of aches, pains or other physical problems			30		48	–14	31	
9	Thought of hurting myself	29	58	58					
10	Talking to people has felt too much for me	65	–5	61	✓			62	
11	Tension and anxiety have prevented me doing important things	69	–3	65		61	11	67	✓
12	Happy with the things I have done +	77	–13	68		21	63	69	
13	Disturbed by unwanted thoughts and feelings	50	22	61	✓	70	–2	61	
14	Felt like crying	55	22	65		56	14	65	✓
15	Felt panic or terror	51	24	62		80	–13	63	✓
16	Made plans to end my life	15	67	48	✓				
17	Optimistic about my future +	73	17	79		66	22	80	
18	Difficulty getting to sleep or staying asleep	44	15	50		43	16	51	
19	Felt warm and affection for someone	41	–4	36		22	70	36	
20	Problems have been impossible to put to one side	70	7	72	✓	59	24	74	✓
21	Able to do most things I needed to +	62	–8	56				57	
22	Threatened or intimidated another person	3	55	31					
23	Despairing or hopeless	72	18	80		59	29	80	
24	It would be better if I were dead			67					
25	Felt criticised by other people			57		58	1	57	
26	Thought I have no friends	53	14	60				60	
27	Felt unhappy	80	8	83				83	
28	Unwanted images or memories have been distressing me			60		76	–13	59	
29	Irritable when with other people			58		51	15	58	
30	I am to blame for my problems and difficulties	55	9	59				60	
31	Overwhelmed by my problems	64	–17	53		1	67	54	
32	Achieved the things I wanted to +	77	–15	66	✓	21	62	68	✓
33	Humiliated or shamed by other people			55		62	–6	55	
34	Hurt myself physically or taken dangerous risks with my life	0	71	36					

+ = positively phrased items and are reverse scored.
CORE-OM = Clinical Outcomes in Routine Evaluation-Outcome Measure.
✓ = item forms a part of Mokken scale.

DOI: 10.1002/cpp

Table 3. Mokken scale of all CORE-OM checked for violations of monotone homogeneity and double monotonicity (n = 543)

CORE-OM item (or paraphrase of item)	Mean	H
16. I made plans to end my life	0.24	0.42
10. Talking to people too much for me	1.58	0.42
*7. Able to cope when things go wrong	1.95	0.40
13. Disturbed by unwanted thoughts/feelings	1.96	0.41
20. Problems impossible to put to one side	2.25	0.47
*32. I have achieved the things I wanted to	2.26	0.45

*These items are positively worded and reverse scored.
p = 0.00067; Scale H = 0.43; Rho = 0.78; Mean = 10.24; standard deviation = 4.56; Skewness = −0.14; Kurtosis = −0.69.
CORE-OM = Clinical Outcomes in Routine Evaluation-Outcome Measure.

Table 4. Mokken scale of CORE-OM minus risk items checked for violations of monotone homogeneity and double monotonicity (n = 543)

CORE-OM item (or paraphrase of item)	Mean	H
15. I have felt panic or terror	1.34	0.33
1. I have felt terribly alone and isolated	1.66	0.38
11. Tension/anxiety prevented me doing things	1.88	0.42
*7. Able to cope when things go wrong	1.95	0.41
14. I have felt like crying	2.10	0.37
20. Problems impossible to put to one side	2.25	0.38
*32. I have achieved the things I wanted to	2.26	0.45
2. I have felt tense, anxious or nervous	2.52	0.45

*These items are positively worded and reverse scored.
p = 0.00063; Scale H = 0.40; Rho = 0.86; Mean = 15.97; standard deviation = 6.81; Skewness = −0.20; Kurtosis = −0.82.
CORE-OM = Clinical Outcomes in Routine Evaluation-Outcome Measure.

or greater, these marker variables are a repeat of the earlier analysis, being item 27 'unhappy' with a component loading of 0.83, item 23 'despairing or hopeless' (0.80) and item 17 'overwhelmed by my problems' (0.80).

Mokken Scale Analysis

The scale derived from the CORE-OM non-risk items is shown in Table 4. It is composed of eight items and has acceptable scaling properties (H = 0.40), reliability (Rho = 0.86) and probability (p = 0.00063). This scale covers a range of items, in terms of difficulty, running from feeling 'tense, anxious or nervous' (item 2) at the less distressed end of the scale to feeling 'panic or terror' (item 15) at the other. (See Tables 3 and 4.)

DISCUSSION

With all 34 Items Included

When the 34 CORE-OM items were assigned to the four domains, as intended by the test authors, the resulting Cronbach's alpha coefficients were never less than 0.75 and reached 0.87 and 0.89. The latter figures, for 'Functioning' and 'Problems or Symptoms', respectively, are at the utmost end of the range necessary for satisfactory internal consistency (Boyle, 1991), and the necessary statistical requirements could still be maintained with items removed.

A series of PCAs, seeking oblique solutions with the full 34 items, found a four-component solution was not justified and that in a three-component solution, the second and third components were each comprised of only three items, thereby failing to meet the necessary standards of reliability and validity.

A two-component solution of 23 and 5 items, respectively did satisfactorily meet the psychometric criteria, except for the latter's marginal Cronbach's alpha of 0.69. The first component was clearly Psychological distress, while the second component was made up of five of the six 'Risk' domain items. The impractically large number of items in the first component suggested a great deal

DOI: 10.1002/cpp

of item redundancy, as was also indicated by the overlarge Cronbach's alpha of 0.94.

When a one-component solution was obtained, it was found that only one item failed to exceed a loading cut-off point of 0.30. At the other extreme, 27 items loaded 0.50 or above, accounting for 42.2% of the total variance, and had an excessive Cronbach's alpha of 0.95. The marker variables were item 27, 'unhappy' with a component loading of 0.83; item 23, 'despairing or hopeless' (0.80); and item 17, 'overwhelmed by my problems' (0.79). Twenty-seven items would be unnecessarily many for a single dimension and the very high Cronbach's alpha equally seems to indicate yet again a redundancy of items.

The Mokken scaling of the CORE-OM inventory with the risk items included produced a very short scale (six items) from the total pool of 34 items. This short scale is anchored at one end in low achievement, running through an inability to ignore problems, disturbing thoughts, an inability to cope, isolation and suicidal ideation at the other end. (This latter being 'I made plans to end my life', the *only* risk item retained in the scale and as such constitutes an example of a 'difficult' item, intended as an indicator of severe depression). Clearly, the inclusion/retention of this 'Risk—harm to self' subdomain item in the CORE-OM proved useful and important in anchoring the Mokken scale. The scale resulting from Mokken scaling not only included a wide range of mean scores on the latent trait (0.24–2.26), but also provided a sensible hierarchy of distress and risk including, and leading to, the ultimate risk of the respondents' planning to take their own lives. This hierarchy is useful as, for example, someone scoring high on being 'disturbed by unwanted thoughts and feelings', but not on the more difficult items (e.g., 'talking to people becoming too much'), is at lower risk than a person scoring high on the latter item. Because of the hierarchical nature of Mokken scales—as explained in the introduction to this paper—a score on the scale will indicate the extent to which the latent trait is present and, thereby, the level of risk. In addition, it also provides a specific descriptor related to that level of self-harm.

With only 28 Non-Risk Items

Turning to the analysis of the 28 CORE-OM items ('Risk' domain items being excluded), this resolved itself into a two oblique component solution of 14 and 5 items. The first component was again Psychological distress, while the second was now dominated by four 'Functioning' items.

The one-component solution found that 25 of the 28 items had loadings greater than 0.50 and formed a dimension with a very large Cronbach's alpha of 0.94. The marker variables were considered to define Psychological distress. They were item 27, 'unhappy', with a component loading of 0.83; item 23, 'despairing or hopeless' (0.80); and item 17, 'overwhelmed by my problems' (0.80).

The Mokken scaling of the CORE-OM non-risk items produced a scale with more items (8) than the scale with the risk items included (6). Three items, 7 'felt able to cope when things go wrong', 20 'Problems have been impossible to put on one side' and 32 'achieved the things I wanted to' were in common. Many of the items are concerned with the way that the respondent has felt (5 of the 8). The scale is sensibly anchored at the lower end (least difficulty) in being 'tense, anxious or nervous' and at the higher end (most difficulty) in feelings of 'panic or terror'. The spread of mean values on the latent trait is 1.34–2.52.

General Issues

It may be argued that in both sets of the aforementioned PCAs, 'Bedford's (1997) criteria' on item inclusion/exclusion are too lax to obtain clearly distinct and conceptually separate components. For example, Deary, Bedford, and Fowkes (1995) insisted upon at least 0.40 for a main loading and no more than 0.20 for cross-loadings as the criteria for item selection in a similar exercise with a successful outcome. When these criteria are applied to both the 34-item and 28-item (without risk) CORE-OM data, it is only possible to derive a one-component solution (Psychological distress). If, however, a less stringent rule, such as 'no cross-loading shall be 0.30 or greater' is added to the earlier criteria, no change is effected in the 34-item pool. By contrast, for the 28 non-risk items, two items would now violate that more modest new ruling, both effecting the five-item second component. Their exclusion would invalidate the two-component solution.

We have deliberately presented the PCAs in full detail in order to emphasize the attempts made to find a viable structure for such a widely used item pool. To state the obvious, it depends upon how much tolerance is allowed for cross-loadings, and therefore, the degree of correlation between components. The greater the tolerance, the less clearly

DOI: 10.1002/cpp

defined is the conceptual nature of the components. Nevertheless, like Lyne et al. (2006), we failed to replicate the supposed domains and subdomains claimed for the CORE-OM.

Within Foulds' personal illness theory (Foulds, 1965), we excluded the 'Risk' domain items as a minimum requirement. A similar case as indicated earlier, can now be made, based on empirical evidence, for the 'Functioning' items likewise to be excluded as they belong logically to a different universe of discourse.

Measures of these concepts, whether symptom-state or trait-attitude, necessarily have introductory instructions, which emphasize the time scale to be considered by respondents, e.g., 'during the last fortnight', in contradistinction with, e.g., 'usually'. Considering first commonly used British measures of psychological distress, various forms of the GHQ (Goldberg & Williams, 1988) specifically enquire about change, e.g., 'no more . . .', 'rather more . . .' or 'much more than usual'. The introductory instructions ask '. . . how your health has been in general, *over the past few weeks*' (italics original), while later adding '. . . we want to know about present and recent complaints, not those that you had in the past.'. Another symptom-state measure, the notebook version of the Hospital Anxiety and Depression Scale (Zigmond & Snaith, 1983), is concerned with how respondents '. . . have been feeling **during the previous week**.' (bold original), specifically stating that this fact should be verbally highlighted. Within the psychiatric field, a major structured interview, the '*Present State* Examination' (italics added) (Wing, Cooper, & Sartorius, 1974), by its very name, emphasizes the relevant time scale being considered.

Second, with regard to the widely used psychometric measures of personality traits, such as the various Eysenck scales and the NEO-FFI (Costa & McCrae, 1992), while not mentioning a specific time scale in their questionnaire's instructions, the booklets' questions or possible responses use language with clear temporal implications. The EPI (Eysenck & Eysenck, 1964) questions ask respondents 'Do you . . .' or 'Are you . . .' 'always', 'often', 'usually', 'tend', 'sometimes' and 'ever'. Similarly, the NEO uses the terms 'always', 'generally', 'often', 'sometimes', 'seldom', 'rarely' and 'never'. Implicit in the relationship between these two questionnaires' statements and the above responses is constancy in the individual's behaviours.

There are two further issues that may be considered in this examination of the CORE-OM's items. The present CORE-OM domains could be individually represented as being the equivalent of other psychological concepts, namely Subjective Well-Being (Happiness), Symptoms or Problems (Psychological distress), Functioning (Current social and interpersonal relationships) and Risk (Intropunitiveness).

Elsewhere the problem has been raised as to the optimum/minimum number of items required for a 'symptom' scale as opposed to a personality trait dimension (Bedford, Grant, dePauw, & Deary, 1999, p. 260) and we have here presented the general view of Kline (1993). It might prove fruitful if this issue was publicly resolved with, ideally, the provision of guidelines.

When the Mokken scaling technique findings are considered in relation to the results of the two-component solution, it will be noted that of the six items in the Mokken scale, five also form a part of the first principal component, i.e., Psychological distress. Likewise, when only 28 CORE-OM items are being considered, 5 of the 8 Mokken scale items form a part of that analysis's Psychological distress component. The two Mokken scales differ conceptually in that the former includes a depression item while the latter includes several anxiety symptoms. (See Tables 3 and 4.) When the three shared items (7, 20 and 32) are excluded, the Mokken scales correlate 0.68 ($p < 0.001$).

The demonstration of hierarchies of items with the CORE-OM (with and without the risk items) increases its utility and improves our theoretical understanding of the underlying construct being measured. In terms of utility, the demonstration of the relationship between specific items and a relative level of distress increases the value of the items as markers of distress. In terms of theory, the relationship between levels of distress, as exemplified by items in the CORE-OM inventory, increases our understanding of the latent trait, its components and its development.

CONCLUSIONS

The intention of the CORE-OM creators was to attempt to not just create a symptom measure, but additionally to incorporate one relating to interpersonal relationships and another concerning physical violence to oneself or others. In the current report it was argued that this was theoretically unsound, as it compounded different universes of discourse, which should not be summated.

DOI: 10.1002/cpp

Empirically, this is the second large UK NHS database, of the responses of adults seeking psychological therapy, which has failed to confirm the suggested structure of domains and subdomains. By contrast, no detailed confirmatory evidence other than the less detailed and explicit report of the questionnaire's constructors has been published. What the two studies do find in common is that the 34 CORE-OM items, when subjected to PCA with oblique rotation, replicating the test authors' procedure, reduce to a large Psychological distress component and a much smaller risk/intropunitive component. The former could have its items drastically reduced and even be resolved into Form A and Form B (and perhaps even Form C?) while still adequately meeting psychometric statistical requirements.

In a similar context, we are reminded of Kline's (1993, p. 558) sardonic and mock-pessimistic stricture on the need to limit the number of items in a questionnaire: 'The fact that the GHQ-12 seems to work virtually as well as GHQ-60 suggests that it may well be that one question such as "Do you feel awful?" might be as effective as the test . . .'. In terms of the MSP analysis, the reduction was a much more modest one, being from 34 and 28 to 6 and 8 CORE-OM items, respectively. The results of PCAs suggest the possibility of a drastic reduction in the psychological distress item pool into a scale of, say, a more economical 10–12 items with varied item content. Additionally, a separate very brief scale of either 'risk' or 'functioning' could be used depending upon the requirements of the test user.

Somewhat on these lines, the test constructors themselves, in addition to producing 11 versions of the CORE-OM in other languages with a further four in progress, have also produced (or are producing) a range of versions for specific groups. These include versions for general practitioners (14 items), young persons (10 items) and people with learning difficulties (in preparation). Additionally, two short forms of 18 items can be obtained, while there is a CORE-10 and a CORE-5 in preparation. In the construction of the CORE-10, items were presumably chosen to ensure a breadth of item content. In constructing smaller scales, it is clearly difficult or impossible, to obtain the same CORE-OM balance in the number of items between the domains and subdomains while attempting a variety of item content. The notion of domains and subdomains then becomes severely threatened and appears for some purposes to have been abandoned.

ACKNOWLEDGEMENTS

We wish to express our thanks to the clinical and administrative members of the service, the participants and Goran Lukic who has coped with the demands of the co-authors and the database.

REFERENCES

Bedford, A. (1997). On Clark–Watson's tripartite model of anxiety and depression. *Psychological Reports*, *80*, 125–126.

Bedford, A., Grant, E., dePauw, K., & Deary, I.J. (1999). The Personal Disturbance Scale (DSSI/sAD): structural cross-validation and proposed short forms. *Personality and Individual Differences*, *27*, 251–261.

Boschen, M.J., & Oei, T.P.S. (2006). Factor structure of the Mood and Anxiety Symptom Questionnaire does not generalize to an anxious/depressed sample. *Australian and New Zealand Journal of Psychiatry*, *40*, 1016–1024.

Boyle, J.G. (1991). Does item homogeneity indicate internal consistency or item redundancy in psychometric scales? *Personality and Individual Differences*, *12*, 291–294.

Caine, T.M., Foulds, G.A., & Hope, K. (1967). *Manual of the Hostility and Direction of Hostility questionnaire*. London: University of London Press.

Cattell, R.B. (1966). The scree test for the number of factors. *Multivariate Behavioural Research*, *1*, 141–161.

Clark, L.A., & Watson, D. (1991). Tripartite model of anxiety and depression: psychometric evidence and taxonomic implications. *Journal of Abnormal Psychology*, *100*, 316–336.

Cortina, J.M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*, 98–104.

Costa, P.T., & McCrae, R.R. (1992). *NEO PI-R professional manual*. Odessa, FL: Psychological Assessment Resources, Inc.

Deary, I.J., Bedford, A., & Fowkes, F.G.R. (1995). The Personality Deviance Scales: Their development, associations, factor structure and restructuring. *Personality and Individual Differences*, *19*, 275–291.

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

Evans, C.E., Cornell, 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.

Eysenck, H.J., & Eysenck, S.B.G. (1964). *Manual of the Eysenck Personality Inventory*. London: University of London Press.

Foulds, G.A. (1965). *Personality and personal illness*. London: Tavistock Press.

Foulds, G.A. (1971). Personality deviance and personal symptomatology. *Psychological Medicine*, *71*(1), 222–233.

Foulds, G.A. (1976). *The hierarchical nature of personal illness*. London: John Wiley.

DOI: 10.1002/cpp

Goldberg, D., & Williams, P. (1988). *A user's guide to the general health questionnaire*. Windsor, UK: NFER-Nelson.

Hulin, C.L., Drasgow, F., & Parsons, C.K. (1983). Introduction to item response theory. In C.L. Hulin, F. Drasgow, & C.K. Parsons (Eds), *Item response theory* (pp. 13–74). Homewood, IL: Dow Jones-Irwin.

Keogh, E., & Reidy, J. (2000). Exploring the factor structure of the Mood and Anxiety Symptom Questionnaire. *Journal of Personality Assessment*, *74*, 106–125.

Kline, P. (1993). *The handbook of psychological testing*. London: Routledge.

Kline, P. (1994). *An easy guide to factor analysis*. London: Routledge.

Lyne, K.J., Barrett, P., Evans, C., & Barkham, M. (2006). Dimensions of variation on the CORE-OM. *British Journal of Clinical Psychology*, *45*, 185–203.

Meijer, R.R., Sijtsma, K., & Smid, N.G. (1990). Theoretical and empirical comparison of the Mokken and Rasch approach to IRT. *Applied Psychological Measurement*, *3*, 283–298.

Mokken, R.J., & Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement*, *6*, 417–430.

Molenaar, I.W., & Sijtsma, K. (2000). *MSP5 for Windows*. Groningen, the Netherlands: iec ProGAMMA.

Sijtsma, K., Debets, P., & Molenaar, I.W. (1990). Mokken scaling analysis for polychotomous items: theory, a computer programme and an empirical application. *Quality & Quantity*, *24*, 171–188.

Singleton, N., Bumpstead, R., O'Brien, M., Lee, A., & Meltzer, H. (2000). *Psychiatric Morbidity among adults living in private households*. London: Stationery Office.

van Schur, W.H. (2003). Mokken scale analysis: between the Guttman scale and parametric item response theory. *Political Analysis*, *11*, 139–163.

Watson, R., Deary, I., & Austin, E. (2007). Are personality trait items reliably more or less 'difficult'? Mokken scaling of the NEO-FFI Personality and. *Individual Differences*, *43*, 1460–1469.

Watson, R., Deary, I., & Shipley, B. (2008a). A hierarchy of distress: Mokken scaling of the GHQ-30. *Psychological Medicine*, *28*, 575–579.

Watson, R., Roberts, B., Gow, A., & Deary, I. (2008b). A hierarchy of items within Eysenck's EPI. *Personality and Individual Differences*, *45*, 333–335.

Wing, J.K., Cooper, J.E., & Sartorius, N. (1974). *The measurement and classification of psychiatric symptoms*. London: Cambridge University Press.

Zigmond, A.B., & Snaith, R.P. (1983). The Hospital Anxiety and Depression Scale. *Acta Psychiatrica Scandinavica*, *67*, 361–370.

DOI: 10.1002/cpp