Title: Measuring patients' views: a bi-factor model of distinct patient-reported outcomes in psychosis

Short title: Measuring distinct patient-reported outcomes in psychosis

Authors:
Ulrich Reininghaus[a,*], Rosemarie McCabe[a], Tom Burns[b], Tim Croudace[c], Stefan Priebe[a]

[a] Queen Mary University of London, Unit for Social and Community Psychiatry, Barts and the London School of Medicine, London, UK
[b] University Department of Psychiatry, Warneford Hospital, Oxford, UK
[c] Department of Psychiatry, University of Cambridge, UK

Correspondence to:
Ulrich Reininghaus, Queen Mary University of London, Unit for Social and Community Psychiatry, Barts and the London School of Medicine
Postal address: Newham Centre for Mental Health, London E13 8SP, United Kingdom; e-mail: u.reininghaus@qmul.ac.uk

Word count (Abstract): 234

Word count (Text): 4,447

Abstract

Background: Patient-reported outcomes (PROs) are widely used for evaluating the care of patients with psychosis. Previous studies have reported a considerable overlap in the information captured by measures designed to assess different outcomes. This may impair the validity of PROs and makes an a priori choice of the most appropriate measure difficult when assessing treatment benefits for patients. We aimed to investigate the extent to which four widely established PROs (subjective quality of life (SQOL), needs for care, treatment satisfaction, and the therapeutic relationship) provide distinct information independent from this overlap.

Methods: Analyses, based on item response modelling, were conducted on measures of SQOL, needs for care, treatment satisfaction, and the therapeutic relationship in two large samples of patients with psychosis.

Results: In both samples, a bi-factor model matched the data best, suggesting sufficiently strong concept factors to allow for four distinct PRO scales. These were independent from overlap across measures due to a general appraisal tendency of patients for positive or negative ratings and shared domain content. The overlap partially impaired the ability of items to discriminate precisely between patients from lower and higher PRO levels. We found that widely used sum scores were strongly affected by the general appraisal tendency.

Conclusions: Four widely established PROs can provide distinct information independent from overlap across measures. The findings may inform the use and further development of PROs in the evaluation of treatments for psychosis.

Introduction

Patient-reported outcomes (PROs) have become increasingly important in the evaluation of treatment for patients with psychosis. A PRO can be defined as 'any report coming directly from patients (i.e. study subjects) about a health condition and its treatment' (FDA, 2006). PRO measures can be used to assess the impact of an intervention on one or more aspects of patients' health status, hereafter referred to as PRO concepts. The term 'PRO' has been used in an increasingly inclusive way, referring not only to purely symptomatic outcomes but also to more complex multidomain concepts such as subjective quality of life (SQOL), needs for care, treatment satisfaction, or the quality of the therapeutic relationship. For measures of multidomain concepts, a conceptual framework is generally used, in which items (e.g. satisfaction with physical health) are grouped within domains (e.g. health), and domains within more general PRO concepts (e.g. SQOL). Research evaluating treatment benefits for patients with psychosis has extensively drawn on PROs (McCabe *et al.*, 2007). Regulatory agencies have also proposed including well-validated PROs as effectiveness endpoints in randomised controlled trials (FDA, 2006; EMEA, 2005). Presently, in the UK, service providers are expected to use PROs for assessing the quality of routine care (DH, 2008, 2009).

When assessing treatment benefits for patients through patient-reported measures several distinct outcomes often seem to be relevant. However, using several measures at the same time raises the problem of multiple statistical testing and is associated with an increased burden to respondents and higher study costs. An explicit and theoretically informed choice of which outcome measures are most appropriate to the evaluation of a specific intervention is therefore required (Moher *et al*, 2001; Altman *et al.*, 2001). Empirically, previous studies have reported a considerable overlap of measures designed to assess different outcomes. In these reports, PROs were highly correlated and a single general factor explained more than half of the variance in SQOL, needs for care, treatment satisfaction, and self-rated symptom scores (Priebe *et al.*, 1998; Fakhoury *et al.*, 2002; Hansson *et al.*, 2007). This general factor has been interpreted as a 'general appraisal tendency' (hereafter 'GAT') of patients for positive or negative ratings across measures. Findings may, however, also reflect an overlap in specific life or care domains such as the patients' health, living situation, or the accessibility of services (hereafter 'shared domain content') at a lower level of generality than established PRO concepts (Floyd & Widaman, 1995; Salvi *et al.*, 2005). A high degree of overlap in responses to items may considerably impair the ability of each measure to capture distinct information, which in psychometrics is referred to as 'discriminant validity' (Campbell & Fiske, 1959). It may also affect the extent to which PRO scores are an adequate reflection of the dimensionality of the concept to be measured, commonly referred to as 'structural validity' (Mokkink *et al.*, 2006). More generally, this overlap makes an explicit and theoretically informed choice of the most appropriate measure difficult when evaluating specific interventions for patients with psychosis.

Previous research into the overlap of PROs has been methodologically limited, e.g. by examining the overlap as accounting for covariance among sum scores rather than item responses, making it difficult to draw accurate conclusions. Previous reports have also failed to assess the extent to which established measures still may capture specific variance independent from this overlap. A better understanding of the distinct drivers of covariance among item responses would help increase the discriminant and structural validity of established PROs and justify their inclusion in treatment evaluations. Indeed, it is only recently that increasing attention has been paid to the role of a bifactor model in resolving dimensionality issues in health outcome measurement (Gibbons & Hedeker, 1992; Reise *et al.*, 2007; Gibbons *et al.*, 2007, 2008; Yang *et al.*, 2009). This bifactor model recognizes that patients' responses to an item depend both on a single general factor that explains covariance among all item responses, and also, independently, on specific factors that only account for responses to items of particular life or care domains. This statistical property appears to be particularly relevant in complex measurement situations when "broad" concepts with content heterogeneous items are to be measured (Reise *et al.*, 2007). In the context of assessing multiple correlated PROs in psychosis, the bifactor structure provides an opportunity to disentangle concept-specific variance from overlap due to a GAT and/or shared domain content.

Against this background, we set out to investigate the extent to which four widely established PROs (SQOL, needs for care, treatment satisfaction, and the therapeutic relationship) provide distinct information in patients with psychosis independent from overlap across measures. Specifically, we aimed to examine whether the overlap in the information provided by different measures: (i) allows for formation of distinct PRO scales, which discriminate precisely between patients from lower and higher levels of each PRO (discriminant validity); (ii) affects the extent to which previously proposed PRO scores are an adequate reflection of the dimensionality of the concept to be measured (structural validity).

Method

Participants

The samples were taken from two multi-centre randomised controlled trials, the UK700 (Burns *et al.*, 1999) and DIALOG (Priebe *et al.*, 2007) studies. Patients in the UK700 sample (n = 708) were between 18 and 65 years old (mean = 38.3, S.D. = 11.6), predominantly male (n = 404, 57.1%), and mostly unemployed (n = 629, 88.8%). They were recruited between February, 1994, and April, 1996, from four UK inner-city mental health services in London and Manchester. Most patients had a diagnosis of schizophrenia (n = 270, 38.1%) or schizoaffective disorders (n = 345, 48.7%). As in the

UK700 sample, patients in DIALOG (n = 507) were between 18 and 65 years old (mean = 42.2, S.D. = 11.5), predominantly male (n = 336, 66.3%), and mostly unemployed (n = 427, 84.2%). The DIALOG sample was recruited between December, 2002, and May, 2005, from community psychiatric services in Granada (Spain), Groningen (Netherlands), London (UK), Lund (Sweden), Mannheim (Germany) and Zurich (Switzerland) covering urban and mixed urban–rural areas. DIALOG patients were mostly diagnosed with schizophrenia (n = 354, 69.8%). The median length of illness in years was slightly higher in the DIALOG (median = 14, IQR = 7 to 23) than the UK700 sample (median = 10, IQR = 5 to 18). The data presented here are the assessments made at baseline in both intervention and control arms. More detailed information on the UK700 and DIALOG studies is available in Burns *et al.* (1999) and Priebe *et al.* (2007).

PRO measures

SQOL was measured using the Lancashire Quality of Life Profile (LQOLP) (Oliver *et al.*, 1997) in the UK700 sample, and its short version, the Manchester Short Assessment of Quality of Life (MANSA) (Priebe *et al.*, 1999) in the DIALOG sample. The LQOLP was based on Lehman's approach, operationalising SQOL as satisfaction with life in general and in major life domains (Lehman, 1996). LQOLP and MANSA contain 24 and 12 items, respectively, asking patients to rate their satisfaction with life in general and several life domains on a Likert-type scale from 'couldn't be worse' (rating of 1) to 'couldn't be better' (rating of 7). Priebe *et al.* (1999) reported good convergent validity for the LQOLP and MANSA.

The number of unmet needs for care was assessed using the Camberwell Assessment of Need, patient-rated version (Phelan *et al.*, 1995), in both samples. The CAN assesses health and social needs across 22 domains. Each domain is rated on a 3-point scale distinguishing between 'no need' (rating of 0), 'met need' (rating of 1) and 'unmet need' (rating of 2). Unmet needs for care ratings were reversed coded to achieve consistency in the coding direction across all PROs.

In the UK700 sample, treatment satisfaction was measured using the Patient Satisfaction Questionnaire (PSQ) (Tyrer & Remington, 1979). The PSQ asks patients to rate nine care domains of satisfaction with services each on a four-point scale (ranging from 1 to 4). The Client Satisfaction Questionnaire (CSQ) (Nguyen *et al.*, 1983) was used for assessing treatment satisfaction in the DIALOG sample. The CSQ consists of eight items rated from 1 to 4 (with higher scores indicating greater treatment satisfaction).

A measure of the therapeutic relationship, the Helping Alliance Scale (HAS, Priebe & Gruyters, 1993), was included only in the DIALOG sample. The HAS comprises 5 items rated on a visual analogue scale ranging from 0 ('not at all') to 10 ('extremely well').

Statistical analysis

Parameter estimation and model fit

To examine the dimensionality of the four PROs, analyses, based on item response modelling, were performed using statistical methods appropriate for ordinal item responses. Model estimation used the robust weighted least squares means and variance adjusted (WLSMV) estimator in MPlus, Version 5.2 (Muthén & Muthén, 1998-2009). The WLSMV estimator has been found to be robust to violations of the assumption of underlying normality and to provide asymptotically unbiased modified standard errors for examining model fit (Flora & Curran, 2004). It returns coefficients from a probit-probit item factor model equivalent to the two parameter normal ogive item response theory (IRT) model extended to polytomous items.

The overall model fit of the latent variable models was assessed by computing the root mean square error of approximation (RMSEA; Steiger, 1990), Comparative Fit Index (CFI; Bentler, 1990), and Tucker Lewis Index (TLI; Tucker & Lewis, 1973). A good model fit is generally indicated by a low RMSEA (below .10 for acceptable and below .05 for very good fit; Browne & Cudeck, 1993) and a high Comparative Fit Index (CFI) and Tucker Lewis Index (TLI) (above .90 for acceptable and above .95 for very good fit; Bentler, 1990; Muthén, 1989).

Model building

Path diagrams of the five alternative latent variable models that were estimated to examine the extent to which different PROs provide distinct information are shown in Figure 1. Model 1 denotes a unidimensional model with the general factor explaining covariance among all item responses, which can be interpreted as a GAT of patients for positive or negative ratings across measures. Model 2 is a multidimensional model with distinct but correlated concept factors for each PRO scale, i.e. SQOL, needs for care, treatment satisfaction, and the therapeutic relationship. Model 3 refers to a bifactor model with a general factor independent from uncorrelated concept factors. Model 4 represents a bifactor model with one general factor and several uncorrelated domain factors to account for shared domain content across measures. Model 5 denotes a bifactor model with a general factor, uncorrelated concept factors, and several uncorrelated domain factors (Reise *et al.*, 2007). Factors in the bifactor models were specified as uncorrelated to assess the independence of distinct concept vs. general and domain factors. Domain factors included into Model 4 and 5 were specified as equivalent as possible across the two study samples for LQOLP, MANSA and CAN. More specifically, domain factors on health (D1), socio-economic status (D2), leisure (D3), living situation (D4), friends and intimate relationships (D5), and safety (D6) were included for LQOLP, MANSA and CAN in both samples. Domain factors on religion (D8) and family (D9) were included in Model 4 and 5 in the UK700 sample only to account for domains specific to the LQOLP that were not covered by more

than one item in the MANSA. Additionally, a domain factor on accessibility of services (D7) was specified for shared domain content of PSQ and LQOLP in the UK700 sample and, similarly, for shared domain content of CSQ and HAS in the DIALOG sample.

[Insert Figure 1 about here]

Model comparison tests

The five alternative latent variable models were compared on the basis of model fit of each model to the sample data, magnitude of factor loadings, and scale information functions. Comparison of model fit indices across models was aimed at testing: first, whether there was any overlap in the information across measures as represented by general and domain factors (Model 2 vs. Model 1, 3, 4, or 5); second, whether the overlap was only due to a GAT or additionally accounted for by shared domain content (Model 1 and 3 vs. 4 and 5); and third, whether there were concept factors independent from the overlap to allow for formation of distinct PRO scales (Model 1 and 4 vs. 2, 3, and 5). These comparisons were probed further in a sensitivity analysis to investigate whether *each* PRO had a *sufficiently strong* concept factor by comparing Models 2, 3, and 5 to reduced models (Model 2r, 3r, and 5r, respectively), in which each PRO concept factor was omitted in turn. We used $\Delta\chi^2$-tests to assess whether Models 2, 3, and 5 better matched the sample data than the reduced models.

Factor loadings were computed to investigate the ability of items to discriminate between patients from lower and higher PRO levels (Reise *et al.*, 2007). Total scale information functions, defined as the inverse of measurement error, were calculated based on standard item response Fisher information formulae to assess measurement precision across the full range of each PRO scale (Embretson & Reise, 2000).

Finally, total scores were computed according to the published version of each PRO measure. These were then regressed on the respective latent concept factors adjusted for the general factor in the best fitting model using structural equation modelling. This step in the analysis aimed to assess the structural validity of previously proposed scoring methods, i.e. the extent to which these simpler scores are an adequate reflection of the dimensionality of the concept to be measured (Mokkink *et al.*, 2006).

Results

Descriptive statistics and correlations

Descriptive statistics and correlations of total PRO scores in the UK700 and DIALOG sample are summarised in Table 1. The mean, S.D., and range of total PRO scores were largely similar across samples. There were highly significant correlations of weak to moderate magnitude among total PRO scores.

[Insert Table 1 about here]

Formation of PRO scales independent from overlap

As can be seen in Table 2, a poor model fit was found for Model 1 with one general factor in the UK700 sample. Fit was also poor for Model 2, 3, and 4. By comparison, a bifactor model with one general factor, three concept and nine domain factors provided a good model fit (Model 5). A similar pattern was evident in the DIALOG sample, in which the best fitting model was also Model 5. This model matched the sample data better than a unidimensional model (Model 1), a multidimensional model with four correlated concept factors (Model 2), a bifactor model with one general and four concept factors (Model 3), and a bi-factor model with one general and seven domain factors (Model 4). Sensitivity analyses showed that inclusion of each PRO concept factor significantly improved model fit.

In both samples there were common features to the model results. There was overlap in the information provided by the different measures (Model 2 vs. 5), and there were sufficiently strong concept factors to allow for formation of distinct PRO scales (Model 1 and 4 vs. 5). In addition, we found that the overlap was due not only to a GAT but also shared domain content (Model 1 and 3 vs. 5).

[Insert Table 2 about here]

Impact of overlap on discriminative ability of items

Factor loadings of Model 5 in the UK700 and DIALOG sample are summarised in Table 3 and 4, respectively. For most PSQ, CSQ and HAS items, factor loadings of $\lambda \geq .35$ were observed for the concept factors. For a large number of SQOL and needs for care items, we found factor loadings of $\lambda \leq .35$ indicating that the ability to discriminate between patients from lower and higher PRO levels was markedly impaired by the overlap.

In both samples, factor loadings of $\lambda \geq .35$ were found for more than half of the SQOL and needs for care items on the domain factors of leisure (D3), living situation (D4), and friends and intimate relationships (D5). This was also observed for the domain factors of religion (D8) and family (D9) in the UK700 sample and socio-economic status (D2), safety (D3) and accessibility of services (D7) in

the DIALOG sample. In the UK700 sample, less than half of the items loaded ≥ .35 on the general factor. However, most items loaded ≥ .35 on this factor in the DIALOG sample. Those items loading ≥ .35 on the general and/or domain factors in addition to the concept factor, were largely as discriminating on the concept as on the general or domain factors, with only a few items being more than twice as informative on general, concept, or domain factors (i.e. UK700: LQOLP13, LQOLP15, CAN20; DIALOG: CAN6).

[Insert Table 3 and 4 about here]

Measurement precision after adjustment for overlap

Scale information functions for concept factors in Model 5 are shown in Figure 2. In both samples, there was a concentration of information coverage at particular points of PRO scales. Patients in the UK700 sample could be precisely scaled around the mean only of the SQOL scale. Information coverage was largely low for needs for care and treatment satisfaction factors and concentrated around the mean in this sample. By comparison, information coverage was mostly higher for concept factors in the DIALOG than UK700 sample. There was a concentration of information coverage in the more positive range of PRO scales in the DIALOG sample.

[Insert Figure 2 about here]

Associations between latent factors and sum scores

Findings on the relationship of latent factors and sum scores are presented in Table 5. While significant associations were observed for distinct concept factors and sum scores whilst controlling for the general factor, only the effect of the treatment satisfaction factor on PSQ sum scores in the UK700 sample was greater than .90. Sum scores were markedly affected by the GAT as indicated by strong associations of sum scores with the general factor.

[Insert Table 5 about here]

Discussion

Main findings

This is the first study to report the extent to which four widely established PROs can provide distinct information in patients with psychosis. Analyses, based on item response modelling, yielded consistent findings in two large samples. First, a bifactor model best matched the data in both samples, suggesting that PROs can be assessed independently from overlap across measures. This

overlap was found to be due both to a GAT and shared domain content. Second, the ability of items to discriminate between patients across PRO levels was largely unaffected by the overlap for measures of treatment satisfaction and the therapeutic relationship. By comparison, SQOL and needs for care items were markedly more impaired in their discriminative ability. Third, findings were complemented by evidence that, after accounting for the GAT and shared domain content, individuals could not be precisely scaled through the full range of each PRO. Lastly, findings on the relationship of latent factors and sum scores suggested that, taking the overlap into account, sum scores did not adequately reflect the dimensionality of the concept to be measured.

Methodological considerations

Patients included in each study were not randomly selected to represent all patients in the given service and were recruited in the context of a RCT. Selection biases might have influenced PRO ratings and findings may not be readily generalisable to all patients with severe and enduring psychosis in routine mental health care.

We sought to determine the replicability of findings by fitting alternative latent variable models in two independent samples (Cudeck & Browne, 1983). However, factor loadings and scale information functions varied to some extent across samples. For the only fully identical measure in both samples, i.e. the CAN, differences in item difficulties, large standard errors, and a better match of item difficulties and latent PRO level may have accounted for the differences in scale information functions. However, 95% confidence intervals still overlapped across the two samples. One may therefore argue that in larger samples estimates of these parameters may have converged (Tsutakawa & Johnson, 1990; Bjorner *et al.*, 2007). Differences in findings may also represent actual differences in the psychometric qualities of the not fully identical measures. For instance, the concentration of information around the mean for the PSQ as compared to high information coverage in the upper range for the CSQ may reflect differences in how precise these measures are (Embretson & Reise, 2000). Based on these findings, one may consider calibrating PSQ and CSQ items into a single scale to increase information coverage across the range of treatment satisfaction levels.

When considered in the context of parsimony, the model that best matched the sample data included more freely estimated parameters than the alternative models considered in this study. In psychometrics, more parsimonious models are commonly considered preferable (James *et al.*, 1982). However, according to the RMSEA, an index sensitive to the number of freely estimated parameters (Steiger, 1990), the bifactor model still matched the sample data best. We would also conclude that it is the conceptual breadth and heterogeneity currently seen in PRO measurement that inevitably requires less parsimonious models. Whether or not this heterogeneity is conceptually justified

remains to be established (i.e. whether definitions of established PROs are sufficiently distinct to warrant them being measured in a single study).

Comparisons with previous research

There is a wealth of research into PROs in psychiatry. Numerous PRO measures have been developed since the late 1970s when they became increasingly relevant to capture the impact of deinstitutionalisation and new psychopharmacological treatments (Kilian & Angermeyer, 1999). The psychometric qualities of PROs in patients with psychosis have, however, rarely been studied considering more than one outcome at a time and using rigorous psychometric methodology. Those studies examining several PROs have consistently found a considerable overlap of measures. These studies emphasised the role of a GAT of patients for positive or negative ratings (Priebe *et al.*, 1998; Fakhoury *et al.*, 2002; Hansson *et al.*, 2007). They were, however, methodologically limited and did not account for half of the, potentially concept-specific, variance that remained unexplained. Also, they identified the problem of the overarching impact of the GAT on different measures without showing a way to advance the methodology of PROs to overcome the problem. Our study has gone a step further. Drawing on recent advances in psychometrics, the bifactor model has provided an approach to consider the GAT and still identify the distinct information provided by four widely established PROs.

Echoing previous reports, we found evidence that PROs are influenced by the GAT. While this tendency needs to be accounted for when assessing distinct outcomes in psychosis, this finding can also be interpreted in the context of recent efforts to reduce multiple outcomes into one overall measure (Leese *et al.*, 2008; Speechley *et al.*, 2009). Based on our findings one can argue for using the general factor as an aggregate PRO, e.g. as a surrogate outcome in the modelling or exploratory phase of evaluating interventions. Our study adds to previous work by showing that, over and above the GAT, there was overlap due to shared domain content (Floyd & Widaman, 1995), which needs to be taken into account when assessing distinct PROs. However, a case can be made that domain factors may provide clinically actionable information at a low level of generality in the evaluation of routine care. Most importantly, however, our findings suggest that established PROs can provide mutually distinct information in patients with psychosis. They conflict with the idea that PROs may solely capture the same underlying concept (Hansson *et al.*, 2007).

There are only a few studies that have examined the discriminative ability of the PRO measures used in the current study. None of the reports that we are aware of has simultaneously accounted for overlap across measures. While we found measures of treatment satisfaction and the therapeutic relationship to be largely unaffected in their discriminative ability, those of SQOL and needs for care were markedly more impaired by the overlap. This may reflect limitations in the conceptual

distinctiveness of these concepts. That is, some concepts have never been conceptually examined as to whether they are sufficiently distinct from already established concepts so that they should be measured independently (Campbell & Fiske, 1959).

There has been even less research into information coverage of PROs in psychosis. Previous studies have almost exclusively reported psychometric properties based on classical test theory (e.g. Oliver *et al.*, 1997; Gaite *et al.*, 2000). We found high information coverage for more favourable PRO levels in one of the psychosis samples. This suggests that evaluation using current PRO measures may provide a more precise picture of positive than of negative patient views (Williams, 1994; Crow *et al.*, 2002; Elwyn *et al.*, 2007; Priebe, 2007). Low information coverage, more generally, makes it difficult to shorten scales whilst maintaining measurement precision through the full range of each PRO (Rodebaugh *et al.*, 2004; Uher *et al.*, 2008). Without shortening scales it is difficult to reduce the assessment burden on patients, which appears to be particularly important in vulnerable patients with psychosis (Gilbody *et al.*, 2002; Gibbons *et al.*, 2008).

Measuring distinct PROs in psychosis

Discriminant and structural validity are highly relevant for determining the value of established measures when assessing PROs in psychosis (Altman *et al.*, 2001; Moher *et al*, 2001; FDA, 2006; Mokkink *et al.*, 2006). Our finding, that using the bifactor model four important PROs can provide distinct information, represents an essential step for establishing discriminant validity of PROs. The measures examined in the current study contain items with high discriminative ability, which can be used in psychosis outcome evaluations. They address different levels of generality (i.e. domains, concepts, and aggregate outcomes). Which of these levels is most useful depends on the purpose of the given evaluation. At present, the structural validity of established measures remains limited, as simple sum scores do not adequately reflect the dimensionality of PROs.

Future research faces the challenge to implement model-based approaches to scoring into outcome evaluations, which can be achieved through item banking and computer-based assessments. Item banks are developed from large pools of items from many available instruments applying item response modelling combined with qualitative methods in an iterative process. This approach has recently been used for developing national item banks for use in research and routine care to improve the measurement of PROs in populations other than psychosis (e.g. Fries *et al.*, 2005). Psychometric research using rigorous methods such as item banking based on a clearly defined conceptual framework of PROs may now be required in psychosis studies.

Overall, findings suggest that advanced analytic methods can help disentangle the complex overlap of PROs. The bi-factor model provides a reasonable explanation of existing data and future studies

measuring more than one PRO may adjust results for the overlap. Different PROs appear to contain distinct as well as shared information, which should be considered in the use and further development of PROs.

Acknowledgements

This work was supported by a Research Training Fellowship funded by the National Institute of Health Research, UK, to U.R. The report is independent research and the views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health.

Declaration of interest

None of the authors reported potential conflicts of interests.

References

Altman D, Schulz K, Moher D, Egger M, Davidoff F, Elbourne D, Gøtzsche P, Lang T (2001) The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Annals of Internal Medicine* **134,** 663-694.

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

Bjorner J, Chang C, Thissen D, Reeve B (2007). Developing tailored instruments: item banking and computerized adaptive assessment. *Quality of Life Research* **16,** 95–108.

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

Burns T, Creed F, Fahy T, Thompson S, Tyrer P, White I, the UK700 group (1999). Intensive versus standard case management for severe psychotic illness: a randomised trial. *Lancet* **353**, 2185-2189.

Campbell D, Fiske D (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin* **56**, 81-105.

Crow, R., Gage, H., Hampson, S., Hart, J., Kimber, A., Storey, L. & Thomas, H. (2002). The measurement of satisfaction with healthcare: implications for practice from a systematic review of the literature. *Health Technology Assessment* **6**(32).

Cudeck R, Browne R (1983). Cross-validation of covariance structures. *Multivariate Behavioral Research* **18**, 147-167.

Department of Health (2008). High quality care for all. NHS next stage review final report. Department of Health: London.

Department of Health (2009). Guidance on the routine collection of Patient Reported Outcome Measures (PROMs). Department of Health: London.

Elwyn G, Buetow S, Hibbard J, Wensing M (2007). Respecting the subjective: quality measurement from the patient's perspective. *British Medical Journal* **335**, 1021-1022.

Embretson S, Reise S (2000). *Item response theory for Psychologists.* Lawrence Erlbaum Associates, Inc.: Mahwah, New Jersey.

EMEA (2005). Reflection paper on the regulatory guidance for the use of health-related quality of life (HRQL) measures in the evaluation of medicinal products.

Fakhoury W, Kaiser W, Roeder-Wanner U, Priebe S (2002). Subjective evaluation: is there more than one criterion? *Schizophrenia Bulletin* **28**, 319-327.

FDA (2006). Guidance for industry. Patient-reported outcome measures: use in medical product development to support labeling claims. US Food & Drug Administration: Rockville, MD.

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

Floyd F, Widaman K (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment* **7**, 286-299.

Fries J, Bruce B, Cella D (2005). The promise of PROMIS: Using item response theory to improve assessment of patient-reported outcomes. *Clinical and Experimental Rheumatology* (Suppl.) **23,** S53-S57.

Gaite L, Vazquez-Barquero J, Arrizabalaga A, Schene A, Welcher B, Thornicroft G, Ruggeri M, Vazquez-Bourgon E, Retuerto M, Leese, M (2000). Quality of life in schizophrenia: development, reliability and internal consistency of the Lancashire Quality of Life Profile - European Version. *British Journal of Psychiatry* (Suppl.) **177**, S49-S54.

Gibbons R, Hedeker D (1992). Full-information item bi-factor analysis. *Psychometrika* **57**, 423-436.

Gibbons R, Bock D, Hedeker D, Weiss D, Segawa E, Bhaumik D, Kupfer D, Frank E, Grochocinski V, Stover A (2007). Full-information bifactor analysis for graded response data. *Applied Psychological Measurement* **31**, 4-19.

Gibbons R, Weiss D, Kupfer D, Frank E, Fagiolini A, Grochocinski V, Bhaumik D, Stover A, Bock R, Immekus J (2008). Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatric Services* **59**, 361-368.

Gilbody S, House A, Sheldon T (2002). Routine administration of Health Related Quality of Life (HRQoL) and needs assessment instruments to improve psychological outcome - a systematic review. *Psychological Medicine* **32**, 1345-1356.

Hansson L, Bjorkman T, Priebe S (2007). Are important patient-rated outcomes in community mental health care explained by only one factor? *Acta Psychiatrica Scandinavica* **116**, 113-118.

James L, Mulai S, Brett J (1982). *Causal analysis: Assumptions, models, and data.* Sage Publications: Beverly Hills, CA.

Kilian R, Angermeyer M (1999). Quality of life in psychiatry as an ethical duty: from the clinical to the societal perspective. *Psychopathology* **32,** 127-134.

Leese M, Schene A, Koeter M, Meijer K, Bindman J, Mazzi M, Puschner B, Burti L, Becker T, Moreno M, Celani D, White I, Thornicroft G (2008). SF-36 scales, and simple sums of scales, were reliable quality-of-life summaries for patients with schizophrenia. *Journal of Clinical Epidemiology* **61**, 588-596.

Lehman A (1996). Measures of quality of life among persons with severe and persistent mental disorders. *Social Psychiatry & Psychiatric Epidemiology* **31**, 78-88.

McCabe R, Saidi M, Priebe S (2007). Patient-reported outcomes in schizophrenia. *British Journal of Psychiatry* (Suppl.) **191**, S21-S28.

Moher D, Schulz KF, Altman DG (2001). The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. *Lancet* **357,** 1191-1194.

Mokkink L, Terwee C, Knol D, Stratford P, Alonso J, Patrick D, Bouter L, de Vet H (2006). Protocol of the COSMIN study: COnsensus-based Standards for the selection of health Measurement INstruments. *BMC Medical Research Methodology* **(6)** 2.

Muthén B (1989). Latent variable modeling in heterogeneous populations. Meetings of Psychometric Society (1989, Los Angeles, California and Leuven, Belgium). *Psychometrika* **54**, 557-585.

Muthén L, Muthén B (1998-2009). *Mplus Version 5.2.* Muthén & Muthén: Los Angeles, CA.

Nguyen T, Attkisson C, Stegner B (1983). Assessment of patient satisfaction: development and refinement of a service evaluation questionnaire. *Evaluation and Program Planning* **6**, 313.

Oliver J, Huxley P, Priebe S, Kaiser W (1997). Measuring the quality of life of severely mentally ill people using the Lancashire Quality of Life Profile. *Social Psychiatry & Psychiatric Epidemiology* **32**, 76-83.

Phelan M, Slade S, Thornicroft G, Dunn G, Holloway F, Wykes T, Strathdee G, Loftus L, McCrone P, Hayward P (1995). The Camberwell Assessment of Need: the validity and reliability of an instrument to assess the needs of people with severe mental illness. *British Journal of Psychiatry* **167**, 589-595.

Priebe S, Gruyters T (1993). The role of the helping alliance in psychiatric community care. A prospective study. *Journal of Nervous & Mental Disease* **181**, 552-557.

Priebe S, Kaiser W, Huxley P, Roder-Wanner U, Rudolf H (1998). Do different subjective evaluation criteria reflect distinct constructs? *Journal of Nervous & Mental Disease* **186**, 385-392.

Priebe S, Huxley P, Knight S, Evans S (1999). Application and results of the Manchester Short Assessment of Quality of Life (MANSA). *International Journal of Social Psychiatry* **45**, 7-12.

Priebe S (2007). Social outcomes in schizophrenia. *British Journal of Psychiatry* (Suppl.) **191**, S15-S20.

Priebe S, McCabe R, Bullenkamp J, Hansson L, Lauber C, Martinez-Leal R, Rossler W, Salize H, Svensson B, Torres-Gonzales F, van den Brink R, Wiersma D, Wright D J. (2007). Structured patient-clinician communication and 1-year outcome in community mental healthcare: cluster randomised controlled trial. *British Journal of Psychiatry* **191**, 420-426.

Reise S, Morizot J, Hays R (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research* **16**, 19-31.

Rodebaugh T, Woods C, Thissen D, Heimberg R, Chambless D, Rapee R (2004). More information from fewer questions: the factor structure and item properties of the original and brief fear of negative evaluation scale. *Psychological Assessment* **16**, 169-181.

Salvi G, Leese M, Slade M (2005). Routine use of mental health outcome assessments: choosing the measure. *British Journal of Psychiatry* **186**, 146-152.

Speechley M, Forchuk C, Hoch J, Jensen E, Wagg J (2009). Deriving a mental health outcome measure using the pooled index: an application to psychiatric consumer-survivors in different housing types. *Health Services and Outcomes Research Methodology* **9**, 133-143.

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

Tsutakawa R, Johnson J (1990). The effect of uncertainty of item parameter estimation on ability estimates. *Psychometrika* **55**, 371–390.

Tucker L, Lewis C (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika* **38**, 1-10.

Tyrer P, Remington M (1979). Controlled comparison of day-hospital and outpatient treatment for neurotic disorders. *Lancet* **313**, 1014-1016.

Uher R, Farmer A, Maier W, Rietschel M, Hauser J, Marusic A, Mors O, Elkin A, Williamson R, Schmael C, Henigsberg N, Perez J, Mendlewicz J, Janzing J, Zobel A, Skibinska M, Kozel D, Stamp A, Bajs M, Placentino A, Barreto M, McGuffin P, Aitchison K (2008). Measuring depression: comparison and integration of three scales in the GENDEP study. *Psychological Medicine* **38**, 289-300.

Williams B (1994). Patient satisfaction: a valid concept. *Social Sciences and Medicine* **38**, 509-516.

Yang F, Tommet D, Jones R (2009). Disparities in self-reported geriatric depressive symptoms due to socio-demographic differences: An extension of the bi-factor item response theory model for use in differential item functioning. *Journal of Psychiatric Research* **43**, 1025-1035.

TABLES

Table 1 Descriptive statistics and mutual correlations of total PRO scores in the UK700 and DIALOG sample

	Mean	S.D.	Min	Max	r (95% CI)		
UK700 sample (n=708)					**LQOLP**	**CAN**	**–**
LQOLP	4.27	.73	1.2	6.5	–	–	–
CAN unmet needs (reversed)	2.64	2.30	0	12	.42 (.49 to .36)	–	–
PSQ	26.1	4.88	9	36	.35 (.27 to .42)	.35 (.42 to .27)	–
DIALOG sample (n=507)					**MANSA**	**CAN**	**CSQ**
MANSA	4.70	.87	2.1	6.9	–	–	–
CAN unmet needs (reversed)	2.86	2.87	0	17	.56 (.50 to .62)	–	–
CSQ	25.7	4.16	8	32	.45 (.38 to .52)	.25 (.16 to .33)	–
HAS	8.0	1.69	.4	10	.37 (.29 to .44)	.16 (.07 to .24)	.61 (.57 to 0.67)

Note: LQOLP, Lancashire Quality of Life Profile; CAN, Camberwell Assessment of Needs; PSQ Patient Satisfaction Questionnaire; MANSA, Manchester Short Assessment of Quality of Life; CSQ, Client Satisfaction Questionnaire; HAS, Helping Alliance Scale.

Table 2 Model fit statistics for unidimensional, multidimensional, and bifactor models in the UK700 and DIALOG sample

	Model 1	Model 2	Model 3	Model 4	Model 5†
UK700 sample					
χ^2	1979.71	1346.38	984.96	929.81	500.90
CFI	.60	.74	.83	.84	.94
TLI	.68	.80	.87	.87	.96
RMSEA	.10	.08	.06	.06	.04
DIALOG sample					
χ^2	1278.82	530.29	643.34	644.13	304.50
CFI	.65	.89	.86	.78	.93
TLI	.74	.82	.90	.82	.94
RMSEA	.11	.06	.07	.09	.05

Note: χ^2, model chi-square; CFI, Comparative Fit Index; TLI, Tucker Lewis Index;
RMSEA, root mean squared error of approximation;
Model 1, unidimensional model with one general factor;
Model 2, multimensional model with correlated concept factors;
Model 3, bifactor model with general and concept factors;
Model 4, bifactor model with general domain factors;
Model 5, bifactor model with general, concept, and domain factors;
† Sensitivity analysis using $\Delta\chi^2$-tests to assess whether Model 5 improved model fit in comparison with Models 5r, i.e. reduced models with each PRO concept factor omitted in turn:

UK700 sample: SQOL ($\Delta\chi^2 = 77.63$, $P < .001$), unmet needs for care ($\Delta\chi^2 = 48.90$, $P < .001$), and treatment satisfaction ($\Delta\chi^2 = 109.82$, $P < .001$);

DIALOG sample: SQOL ($\Delta\chi^2 = 69.79$, $P < .001$), unmet needs for care ($\Delta\chi^2 = 138.34$, $P < .001$), treatment satisfaction ($\Delta\chi^2 = 26.29$, $P < .001$), and therapeutic relationship ($\Delta\chi^2 = 104.68$, $P<.001$).

Table 3. Standardized factor loadings of LQOLP, CAN, and PSQ items in bi-factor model with uncorrelated general, concept, and domain factors (Model 5) in the UK700 sample

Items		Model 5												
		G	**C1**	**C2**	**C3**	**D1**	**D2**	**D3**	**D4**	**D5**	**D6**	**D7**	**D8**	**D9**
LQOLP items														
1	Life as a whole	.53	.21			.10								
2	Job situation	.26	.23				.14							
3	Financial comfort	.30	.07				.83							
4	Money for enjoyment	.38	.07				.81							
5	Getting on with people	.34	.45							.47				
6	Number of friends	.44	.33							.64				
7	Pleasure home acts	.42	.37					.60						
8	Pleasure outside acts	.40	.29					.49						
9	Pleasure radio/TV	.46	.15					.38						
10	Living arrangements	.27	.52						.44					
11	Residence privacy	.23	.42						.53					
12	Continued residence	.25	.50						.42					
13	Residence independence	.31	.36						.74					
14	Residence influence	.34	.42						.62					
15	Other residents	.24	.53						.39					
16	Personal safety	.34	.42								.62			
17	Neighbourhood safety	.33	.42								.71			
18	Family situation	.31	.32											.62
19	Amount family contact	.29	.30											.62
20	Religious faith	.22	.10										.52	
21	Religious practice	.10	.15										.52	
22	Frequency doctor	.43	.33									.24		
23	General health	.39	.38			.41								
24	Mental health	.32	.36			.60								
CAN items														
1	Accommodation	.33		.20					.23					
2	Food	.64		.27					-.05					
3	Looking after the home	.52		.20					-.25					
4	Self care	.32		.29										
5	Daytime activities	.42		.51				.07						
6	Physical health	.37		.09		.39								
7	Psychotic symptoms	.32		.03		.09								
8	Information condition	.35		.24										
9	Psychological distress	.46		.51		.29								
10	Safety to self	.36		.21							.08			
11	Safety to others	.39		-.16							.33			
12	Alcohol	.11		.31										
13	Drugs	.32		-.17										
14	Company	.59		.46						.39				
15	Intimate relationships	.40		.43						.25				
16	Sexual expression	.11		.30						.12				
17	Childcare	.21		-.10										
18	Basic education	.26		.07			-.09							
19	Telephone	.29		.20					-.01					
20	Transport	.10		.49										
21	Money	.44		.22			.15							
22	Benefits	.26		.21			.41							
PSQ items														
1	Ease of access	.35			.33							.57		
2	Appointment times	.31			.44							.66		
3	Time spent with staff	.41			.49							.16		
4	Waiting time	.32			.40							.19		
5	Sensitivity for culture	.31			.54									
6	Knowledge of medication	.38			.56									
7	Decision making	.31			.57									
8	Continuity of care	.28			.60									
9	Links between services	.40			.62									

Note: Model 5, bifactor model with general, concept, and domain factors; G, general factor; C, concept factor; D, domain factor; C1, SQOL; C2, unmet needs for care (reversed); C3, treatment satisfaction; D1, health; D2, socio-economic status; D3, leisure; D4, living situation; D5, friends and intimate relationships; D6, safety; D7, accessibility of services; D8, religion; D9, family.

Table 4. Standardized factor loadings of MANSA, CAN, CSQ, and HAS items in bi-factor model with uncorrelated general, concept, and domain factors (Model 5) in the DIALOG sample

Items		Model 5											
		G	**C1**	**C2**	**C3**	**C4**	**D1**	**D2**	**D3**	**D4**	**D5**	**D6**	**D7**
MANSA items													
1	Life as a whole	.55	.42				.15						
2	Job situation	.37	.48					.21					
3	Financial situation	.41	.19					.47					
4	Friendships	.46	.32								.14		
5	Sex life	.33	.40								.55		
6	Leisure activities	.43	.49						.44				
7	Accommodation	.47	.07							.58			
8	Living situation	.45	.15							.40			
9	Personal safety	.50	.24									.31	
10	Family relationships	.51	.12										
11	Physical health	.36	.18				.54						
12	Mental health	.45	.42				.36						
CAN items													
1	Accommodation	.35		.38						.54			
2	Food	.46		.51						.45			
3	Looking after the home	.37		.21						.36			
4	Self care	.32		.64									
5	Daytime activities	.48		.33					.44				
6	Physical health	.15		.37			.91						
7	Psychotic symptoms	.39		.68			.19						
8	Information condition	.39		.27			.19						
9	Psychological distress	.44		.60									
10	Safety to self	.39		.55								.66	
11	Safety to others	.41		.41								.40	
12	Alcohol	.25		.54									
13	Drugs	.12		.69									
14	Company	.42		.40							.30		
15	Intimate relationships	.37		.23							.75		
16	Sexual expression	.39		.26							.76		
17	Childcare	.10		.38									
18	Basic education	.12		.52				.26					
19	Telephone	.43		.35						.19			
20	Transport	.25		.57									
21	Money	.27		.29				.59					
22	Benefits	.07		.29				.67					
CSQ items													
1	Quality of service	.74			.33								
2	Got service wanted	.67			.54								.35
3	Service met needs	.66			.43								
4	Recommend service to friend	.54			.48								
5	Satisfaction amount of help	.58			.47								
6	Dealing more effectively	.57			.44								
7	Generally satisfied with service	.82			.49								
8	Come back if help needed	.57			.52								
HAS items													
1	Right treatment	.59				.53							.35
2	Understood by therapist	.58				.67							
3	Criticized by therapist	.19				.53							
4	Committed therapist	.43				.79							
5	Trust therapist	.49				.74							

Note: Model 5, bifactor model with general, concept, and domain factors; G, general factor; C, concept factor; D, domain factor; C1, SQOL; C2, unmet needs for care (reversed); C3, treatment satisfaction; C4, therapeutic relationship; D1, health; D2, socio-economic status; D3, leisure; D4, living situation; D5, friends and intimate relationships; D6, safety; D7, accessibility of services.

Table 5. Regression of sum scores on latent factors in best fitting model (Model 5)

	Model 5									
	G		C1		C2		C3		C4	
	β (95% CI)	P	β (95% CI)	P	β (95% CI)	P	β (95% CI)	P	β (95% CI)	P
UK700 sample										
LQOLP	.73 (.64 to .82)	<.001	.77 (.67 to .86)	<.001	–	–	–	–	–	–
CAN unmet needs (reversed)	.65 (.57 to .73)	<.001	–	–	.83 (.75 to .91)	<.001	–	–	–	–
PSQ	.55 (.45 to .65)	<.001	–	–	–	–	.93 (.86 to .99)	<.001	–	–
DIALOG sample										
MANSA	.76 (.69 to .82)	<.001	.79 (.72 to .86)	<.001	–	–	–	–	–	–
CAN unmet needs (reversed)	.54 (.48 to .61)	<.001	–	–	.82 (.74 to .89)	<.001	–	–	–	–
CSQ	.74 (.67 to .80)	<.001	–	–	–	–	.53 (.43 to .63)	<.001	–	–
HAS	.55 (.48 to .62)	<.001	–	–	–	–	–	–	.84 (.79 to .89)	<.001

Note: Model 5, bifactor model with general, concept, and domain factors; G, general factor; C, concept factor; C1, SQOL; C2, unmet needs for care (reversed); C3, treatment satisfaction; C4, therapeutic relationship.

FIGURES

Figure 1. Path diagrams of five alternative latent variable models, compared to examine the extent to which different PROs provide distinct information. Notation: (□) items (observed variables); (○) latent factors (unobserved variables); (→) loadings of items onto latent factors; G, general factor; C, concept factor; D, domain factor; C1, SQOL; C2, unmet needs for care (reversed); C3, treatment satisfaction; C4, therapeutic relationship; Dx, domain factor (example) accounting for shared domain content across measures; Model 1, unidimensional model with one general factor; Model 2, multimensional model with correlated concept factors; Model 3, bifactor model with general and concept factors; Model 4, bifactor model with general and domain factors; Model 5, bifactor model with general, concept, and domain factors.

Figure 1 (cont.)

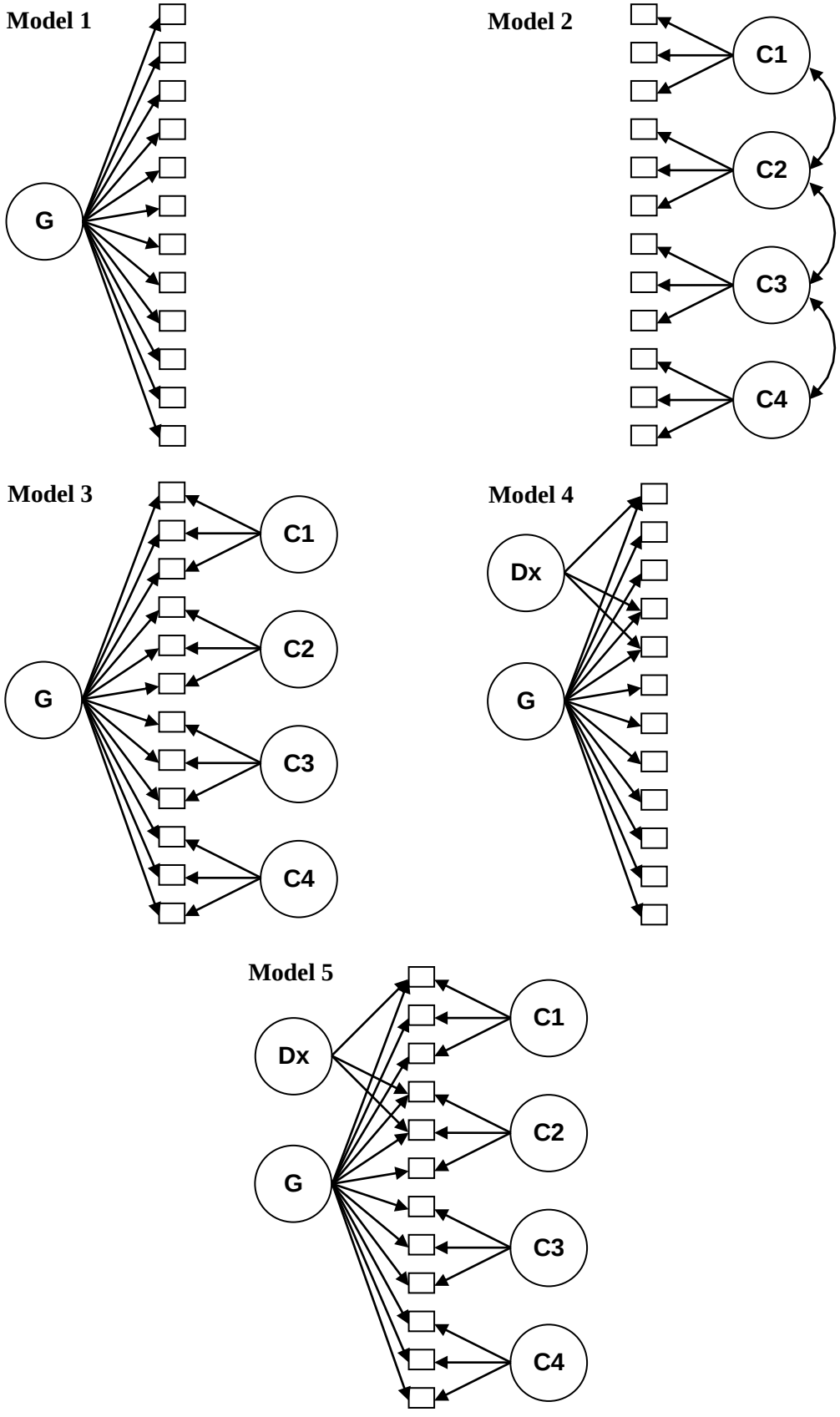

Figure 2. Scale information functions of concept factors in Model 5.

Notation: The line charts represent the scale information function (i.e. the inverse of measurement error, y-axis) across the range of measurement scale from -6 SDs below the mean to +6 SDs above the mean (x-axis); Model 5, bifactor model with general, concept, and domain factors.