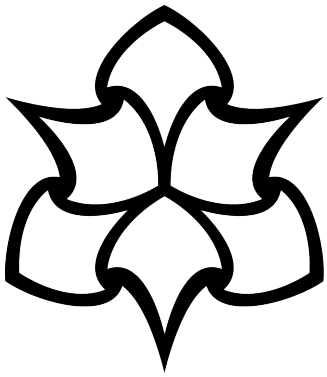

Factor validation and Rasch analysis of the Individual Recovery Outcomes Counter

Running head: Measuring individual mental health recovery

Article category: Original article

Geoffrey L. Dickens[1] Bridey Rudd[1,2] Nutmeg N. Hallett[3] Robin M. Ion[1] Scott M. Hardie[1]

[1]Division of Mental Health Nursing and Counselling, Abertay University, Dundee, United Kingdom

[2]Penumbra, Edinburgh, United Kingdom

[3]School of Health, University of Northampton, Northampton, United Kingdom

Correspondence: Geoffrey L. Dickens, School of Social and Health Sciences, Abertay University, Bell Street, Dundee, DD1 1HG. United Kingdom (T: +44 1382 308257; E: g.dickens@abertay.ac.uk;

Abstract

Objective: The Individual Recovery Outcomes Counter is a 12-item personal recovery self-assessment tool for adults with mental health problems. Although widely used across Scotland, limited research into its psychometric properties has been conducted. We tested its' measurement properties to ascertain the suitability of the tool for continued use in its present form. ***Materials and methods:*** Anonymised data from the assessments of 1,743 adults using mental health services in Scotland were subject to tests based on principles of Rasch measurement theory, principal components analysis and confirmatory factor analysis. ***Results:*** Rasch analysis revealed that the 6-point response structure of the Individual Recovery Outcomes Counter was problematic. Re-scoring on a 4-point scale revealed well-ordered items that measure a single, recovery-related construct, and has acceptable fit statistics. Confirmatory factor analysis supported this. Scale items covered around 75% of the recovery continuum; those individuals least far along the continuum were least well addressed. ***Conclusions:*** A modified tool worked well for many, but not all, service users. The study suggests specific developments are required if the Individual Recovery Outcomes Counter is to maximise its' utility for service users and provide meaningful data for service providers.

Keywords: Mental health, recovery, factor analysis, Rasch measurement theory, validity

Introduction

Recovery, the concept that people can successfully negotiate periods of mental distress or diagnosed mental disorder to lead rich and fulfilling lives, is one of the key drivers of contemporary mental health policy [1,2]. The core recovery construct relates to goals that focus on development of personal meaning rather than on symptom eradication or cure, and it is often conceptualised as a personal and social journey [3,4,5]. Specific elements that have been reported as relevant and important include support from others, finding hope, engaging in meaningful activities, incorporating difficulties with illness or disability, overcoming stigma, taking control, managing mental health symptoms, empowerment, and citizenship [6-8]. The extent to which recovery has become mainstream since the publication of Anthony's [9] seminal paper is evidenced by its embedding in national strategy documents [10,11], and in the presence of multiple literature reviews published in international journals [7,8,12-15].

The growing traction of recovery-informed approaches to service provision has been aided by, and has itself fuelled demand for, the development of instruments that can quantify service user progress towards the goal of personal recovery. However, the psychometric properties of individual measures is variable, and there is little consensus about which are the strongest [16,17]. There is a need to further test existing measures of recovery using rigorous techniques in order to provide evidence for their continued use, or information about how they can be improved.

The Individual Recovery Outcomes Counter (I.ROC), has been developed and implemented by the Scottish mental health charity Penumbra [18]. The 12-item tool is predicated on four hypothetical or notional domains, each comprising three items (see Figure 1). The acronym HOPE is used to capture description of the four domains (Home, Opportunity, People, Empowerment). The tool was developed to i) measure service user outcomes following a period of service contact in order to demonstrate effectiveness; ii)

allow the client or service user to self-monitor progress throughout the service contact; and iii) facilitate therapeutic conversation between service users and support staff by signposting areas key to personal recovery.

Despite its widespread use in Scotland, it is important to clarify that development of the Individual Recovery Outcomes Counter was conducted through extensive consultation with service users and providers rather than via a formal or traditional process of psychometric test development; as a result, the four domains described are notional rather than empirical constructs. The actual empirical research into the psychometric properties of the tool has thus far been limited to a validation study using data collected in its first year of use among *N*=171 adults [11,18]. Principal components analysis [11] revealed a two factor structure (eight *intrapersonal self-reflection/change items* and four *interpersonal outward/forward looking* items) rather than a replication of the notional four domains. There are now considerably more data available, supported by more robust training, data collection methods, and quality control measures, to further assess the tool's properties using two statistical techniques which, primarily due to sample size, could not be conducted previously: Rasch analysis and confirmatory factor analysis.

Traditionally, the testing of the psychometric properties of rating scales in mental health practice has been based on classical test theory [19]. One of the limitations of classical test theory is that, because tests reflect both the ability of the person taking the test (personability) and the difficulty of the test (item difficulty), the information generated is sample specific and cannot readily be used to compare scores across groups or individuals. However, scales based on item response theory, most notably Rasch measurement approaches, are increasingly utilised in test development and validation having previously been relatively neglected due to the larger sample sizes needed compared with tests based on classical test theory [20]. A Rasch model defines how a set of items should perform to

generate reliable and valid measurements. Rasch analysis is used to quantify the extent to which rating scale data (in this case the service user ratings on tool items) fit with predictions of those ratings from the Rasch model [21]. Close fit between the predicted and actual scores indicates valid measurement [22], and thus Rasch analysis can provide detailed diagnostic information about how a scale can be improved, for example through identification of items that fail to fit the Rasch model, and score interpretation [21]. Further, it reveals important information about a tool's ability to test the full range of a single latent trait, in this case 'personal recovery', within the context of use, in this case, people who use the services in which the Individual Recovery Outcomes Counter has been implemented. The underlying assumption of Rasch analysis, therefore, is that the scale under investigation is unidimensional. Further, this assumption is tested during Rasch analysis. Any indication that the scale violates unidimensionality can be resolved by further testing items that appear to comprise separate underlying latent traits as separate unidimensional scales.

In addition to the Rasch analysis, confirmatory factor analysis is also warranted. The previous study of the Individual Recovery Outcomes Counter [23] used principal components analysis, an exploratory procedure, to identify factor structure: i.e. latent variables comprising subgroups of individual items. However, the hypothesis-testing confirmatory procedure may be more appropriate the tool is predicated on a notional four domain structure resulting from considerable user-involvement whose validity as an empirical entity this approach could confirm or refute. It may seem counterintuitive to simultaneously use both factor analysis to explore the potential multidimensionality of the scale and Rasch that assumes that the same scale is univariate. However, we have followed advice that where there is a lack of certainty about the structure of the items under study, and where dimensionality needs to be examined alongside evidence of local dependence, confirmatory factor analysis alone is insufficient, and its use in combination with Rasch is warranted [24,25,26]. More rigorous validation of

the notional four domain structure would allow researchers, practitioners, and service providers to have added confidence in self-assessment data such that results for particular factors, rather than single items, could be interpreted to signify needs for specific targeting.

The aims of the current study therefore were i) to test the extent to which the Individual Recovery Outcomes Counter measures one or more latent recovery-related variables; ii) to provide information about the validity of individual tool items and determine whether and how the the tool might be improved to better measure recovery; iii) to determine whether the notional four domain model is empirically supported.

Materials and Methods

Participants

The study forms part of a larger project examining the development and use of the Individual Recovery Outcomes Counter and involves secondary analysis of an existing but previously unexamined cross-sectional dataset of self-assessment questionnaires completed by clients/ users of mental health and related services in Scotland. All aspects of the study were reviewed and approved by the Abertay University Research Ethics Committee. Eligible participants were adults over 18 years of age who used third sector community mental health services in Scotland, and completed at least one assessment, between January 2012 and October 2014. Sample size was essentially arbitrary and used all data meeting inclusion criteria at the date of collection; however, the final sample exceeded the most conservative guidelines on required absolute sample size [27] and subject-to-variable ratio [28] for use in confirmatory factor analysis. Participants were receiving services from one of 35 projects (*Mdn* participants per service=10, range 1-268) at the time of assessment. Projects included supported living services (k=9); self-harm services (k=5); homeless services and youth projects (both k=2). Characteristics of study participants are presented in Table 1.

>>Insert Table 1 about here<<

Procedure

Anonymised baseline Individual Recovery Outcomes Counter self-assessment and demographic data held on the computerised database were retrieved on 17th November 2014. Scores were cleansed manually to remove duplicate or incomplete iterations.

Measure

The Individual Recovery Outcomes Counter was developed by Penumbra, a Scotland-based mental health charity operating in the third/voluntary sector as a mental health service provider. The tool aims to establish, and subsequently track, an individual's level of personal recovery across four notional domains (*Home, Opportunity, People, Empowerment*). Each domain comprises three items, each described using text (descriptions, related terms) and graphical prompts, an issue important to personal recovery as identified during its extensive development (see Figure 1). For each item, the service user is requested to respond on an ordinal 6-point unipolar scale (1 = *never;* 2 = *almost never*; 3 = *sometimes*; 4 = *often*; 5 = *most of the time*; 6 = *all of the time*) with a higher score intended to represent a greater level of personal recovery. Each rating is intended to refer to the past 3-month period. Scores can be plotted on a radar chart that facilitates visualisation of change over time (see Figure 1). In a preliminary validation study [23], the tool had good internal consistency (α=.86), correlated significantly with the Recovery Scale [29], an established measure of recovery [30], and one of mental health outcome (BASIS-32 [31,32]). Additionally, it was favoured over either of these measures by service users as a personal recovery outcomes measure. The tool has adequate readability (Flesch-Kincaid score 6.2, unpublished data). Previous analysis of a sample of *N*=171 adults revealed that the tool comprised two factors accounting for 51.8% of total variance in scores: an intrapersonal factor largely relating to the individual's inner life (*mental health, life skills, safety and comfort, physical health, personal network, valuing myself, participation and control*, and *self-management*); and an interpersonal factor

comprising four items (*exercise and activity*, *purpose and direction*, *social network*, *hope for the future*) relating to the individual's ability to participate socially, and play a meaningful part both in their own lives and in the wider community [18]. Thus, the notional four domains measured by the tool were not reflected by the results of this prior analysis.

>>Insert Figure 1 about here<<

Training is provided to support workers employed by Penumbra (including author 2) and is mandatory for all those using the tool in practice; further training addresses 'Recovery in Practice', 'Coaching for HOPE', 'Planning for HOPE', and 'Motivational Interviewing'. In practice, the support worker introduces the individual service user to the Individual Recovery Outcomes Counter, and completes a baseline assessment with them during the first four scheduled support sessions. Assessment is then repeated quarterly to support individual work and track change. The assessment questions facilitate an outcomes-focused conversation during which service users can acknowledge progress and identify priorities. Ratings and notes reflecting the conversation are manually entered by the practitioner into a secure online database managed by Penumbra. In addition, information about age, gender, ethnicity, employment status, source of referral, length in service, and reason for leaving service are collected.

The Individual Recovery Outcomes Counter has been used across Penumbra's community based services and within a further 16 third sector organisations since 2012. While it is used most commonly within mental health focused community settings, it is also used within services focusing on related issues including homelessness and substance misuse.

Tests of data quality, distribution, stability, scaling assumptions, reliability, and validity

Data were examined for data quality (percent missing data for each item in excluded data) and for normality of distribution; Hair et al. [28] suggest that data distribution is considered normal if skewness is between -2 and +2 and kurtosis is between -7 and +7. Bond

and Fox [33] suggest that a minimum of 10 responses per scoring category is required to determine an estimation of stable threshold values. Descriptive and correlational analyses were conducted to evaluate scaling assumptions (e.g., similar item mean scores and variances, scores which span the entire measurement continuum, and the magnitude and similarity of corrected item-total correlations). Further analysis was conducted of reliability and validity, scale-to-sample targeting (score means and standard deviation [*SDs*]; floor and ceiling effects), and internal consistency (Cronbach's alpha). Test-retest reliability was measured on a separate sample of *n*=70 staff and students from Abertay University (*M* age = 26.4 years, *SD*=10.7 years, range 18 to 65 years; 73.1% female) who completed the tool twice with an interval of one week. Intraclass correlation coefficients (*r*) were calculated for items and total scores using 2-way mixed models to test the hypothesis that ratings remained stable over the brief interval. A value between 0.75 and 1.00 = *excellent*; 0.60 to 0.74 = *good*; 0.40 to 0.59 = *fair*; and < 0.40 = *poor* [34].

Rasch Measurement Testing of the Individual Recovery Outcomes Counter

Rasch measurement methods were employed to better establish whether the tool captures the full range of the construct of personal recovery in the context of community mental health service use. The Rasch model conceptualizes the measurement scale of a construct as a ruler; a scale that defines the full range of a construct along its whole continuum will comprise scores ranging from ±4 logits (equivalent to ±4 standard deviations). Both test items and test-takers (i.e., service users) can be located along the scale from left to right in terms of their difficulty (less to more) and ability (less to more) respectively (see Figure 2). The Rasch model expresses the likelihood that an item that represents a given level of the construct of interest will correspond with the perceived level of that construct in people with a given level of the construct as a logistic function of the difference between item difficulty and person ability [35].

Rasch measurement provides a choice of two models of parameterization for non-dichotomous data. The rating scale model specifies that a set of items share the same rating scale structure [36] while the partial credit model specifies that each item has its own rating scale structure [37]. Model selection requires consideration of whether the thresholds of the rating scale are known in advance of data collection (not in the case of the Individual Recovery Outcomes Counter), a condition which supports use of the partial credit model [37]. Further, previous research, into recovery-related constructs using secondary data [38-41] have found the amount of partial correctness to vary across items, again supporting use of the partial credit model. We therefore used this model to guide us in establishing whether five important indicators of rigorous measurement were met: fit, targeting, dependency, multidimensionality, and reliability.

Fit. To measure the extent to which items in the Individual Recovery Outcomes Counter work together to capture the individuals' level of personal recovery we tested the performance of each item by visually inspecting for a monotonic ordering of mean item ability, item thresholds, and of the item characteristics curves. Further, to measure the item fit to the Rasch model, the unweighted mean square outfit statistic and the weighted mean square infit statistic were calculated. The outfit statistic is sensitive to unexpected observations by person or item, while infit is sensitive to unexpected patterns where residuals are close to estimated person abilities [42]. Expected values are close to 1.0 with greater values indicating underfit between the items and the model, and lower values indicating overfit (i.e., that the data predict the model too well) and hence item redundancy. Scale validity is more greatly affected by underfit than overfit. Mean square of 0.6–1.4 represents the ideal range [43] items with a value >2.0 are likely to distort or degrade the scale causing inaccurate measurement, while those of 1.4 – 2.0 or <.5 are potentially unproductive for the measurement but not degrading [44]. Finally, we inspected the indices of person and item

reliability and separation which are used to classify people. Low person separation (< 2) or person reliability < 0.8) implies that the instrument may be insufficiently sensitive to distinguish between high and low performers and hence more items may be needed. Item separation and reliability are used to verify the item hierarchy. Low item separation (<3) or item reliability (< 0.9) implies that the person sample is insufficiently large to confirm the item difficulty hierarchy of the instrument; this is equivalent to a measure of construct validity. In the event of inadequate fit then collapsing of categories is recommended [33].

Targeting. We examined how people and items were distributed along the proposed latent personal recovery continuum, and whether the 12 items covered the full range of the continuum and targeted the sample under investigation. This allowed us to gauge the calibration of the instrument to the population by comparing graphically how closely the amount of personal recovery orientation displayed by the respondents was adequately measured by the items on the scale [45]. We also flagged items in similar locations as in need of further investigation because of their potential redundancy.

Dimensionality. Principal component analysis of the residuals is used in Rasch measurement theory to test its underlying assumption that all of the data can be explained by the latent measures (in this instance personal recovery). This differs from the principal components analysis used in classical test theory which is a correlational model that aims to identify factors *within* a scale; principal component analysis of the residuals is a hierarchical implication model where positive responses to difficult items imply positive responses to easy items, but not vice versa [46]. The Rasch model focuses on the unexplained part of the data, the residuals, by extracting the common factor that explains the most residual variance. Following standardisation of each residual, the noise should, if there is no meaningful structure to the residuals and the scale is most likely unidimensional, follow a random normal distribution. Identification of a meaningful structure, indicated by Eigenvalues ≥2, suggests

that the presence of another dimension to the original factor or scale should be investigated [47,48,49].

Item invariance. In order to test whether particular groups of people respond to the scale in a systematically different way to others the Differential Item Functioning (DIF) statistic using the Mantel-Haenszel approach was employed [50]. It could, for example, be predicted that females might respond differently to items such as social network. Testing DIF by gender will facilitate conclusions about whether, and how, this manifests. We tested DIF by gender for all 12 scale items. To analyse DIF, item parameters are held constant while person measures are estimated separately for each group. The effect size, the DIF contrast, is reported in logits and is the difference between the two DIF measures; a substantive DIF is ≥0.64 logits. The statistical significance is computed using t-tests.

Reliability. We assessed reliability using the Person Separation Index [51] which is analogous to the Cronbach's alpha [52]. A value of 0.70 and above is considered acceptable as an indicator for group use, and 0.70 through 0.85 for individual use [51].

Construct validity. Point-measure correlations were calculated to investigate whether all the items within the scale were measuring the same construct. A fundamental concept in Rasch is that higher person measures lead to higher ratings on items and vice versa [44]. The accuracy of this concept is reported by point-measure correlations which should be noticeably positive (>.50). All Rasch analyses were conducted using Winsteps ® 3.81.0 software.

Factor structure of the Individual Recovery Outcomes Counter

Factors comprise multiple variables whose responses are correlated; that covariance is used to infer the presence of latent variables, also known as factors. Factors may have practical value, for instance their presence may suggest interventions that might be targeted at a particular latent trait. Factor analysis facilitates parsimony since only items related to the overall construct and to one constituent factor need be retained in a scale. The potential

importance in the current case is that, should different factors exist, then low scores on a particular factor (e.g., Home items) but not another (e.g., Opportunity items) might help services to target resources at issues which are most problematic and which could have greatest impact on outcome. In addition, the tool is predicated on a hypothesised model based on considerable collaborative development and it is desirable to test that model empirically.

To test for scale-item redundancy Pearson correlations were conducted between all 12 item scores. Correlations between 0.3 and 0.7 indicate that items are sufficiently related to form part of the same latent construct but not so related as to be redundant. Next, a procedure similar to exploratory factor analysis, principal components analysis, was conducted to determine whether the previously reported two factor structure [18] was replicated in this considerably larger sample, or whether the notional four domain structure would now be revealed. To assess the appropriateness of the data for factor analysis, a Kaiser-Mayer-Olkin measure of sampling adequacy was conducted. A score of ≥.90 is described as *excellent* while scores <.50 are *unacceptable* [53,p.58]. To determine the number of factors to be extracted, guidelines described by Costello and Osborne [54] were followed. Multiple analyses were conducted; first by setting the number of factors to be extracted as all those with Eigen values greater than one; second, analyses were run with number of extracted factors manually set i) equal to the number of factors in previously demonstrated analyses (two factors), ii) to the number of factors hypothesised by the tool developers (four factors), and iii) to the number of factors identified from inspection of the 'elbow' on the accompanying scree plot (also two factors). Analyses were also run for the number of factors between one above and one below those numbers. As a result, possible solutions ranging from one to five factors were considered. Oblique rotation was conducted where extracted factors were significantly correlated (>.32), and orthogonal rotation where factor correlation was less evident (<.32; [55]). The most satisfactory factor structure was decided according to i) the smallest number

of cross-loading items, ii) structure comprising factors of three or more items [54], and iii) acceptable internal reliability of factors indicated by Cronbach's alpha. George and Mallery [56,p.231] suggest >.9 = *excellent*; >.8 = *good*; >.7 = *acceptable*; >.6 = *questionable*; >.5 = *poor*; and < .5 = *unacceptable*. Because internal reliability tends to increase with test length [57] we used the Spearman-Brown prophecy formula to calculate the likely internal reliability of each factor scale in the event that it was increased to 10 items of similar quality. Principal components analysis was conducted in IBM SPSS Statistics (V.22.0.0.1).

Finally, confirmatory factor analysis was conducted. While large sample size reduces the problem of multivariate non-normality which might undermine the assumptions of confirmatory factor analysis, we tested this anyway by calculating the Mahalanobis distance in order to identify data outliers and exploring whether they had significant effects on the data structure. Maximum Likelihood Estimation was used to estimate the models' fit using the following indices: Root Mean Square Error of Approximation, the Comparative Fit Index, the Normed Fit Index, the Goodness of Fit Index, and the Adjusted Goodness of Fit Index. For the Root Mean Square Error of Approximation, values < 0.08 and <0.05 reflect reasonable and excellent fits respectively. For the fit indices, values vary along a continuum of 0 to 1 with those >0.9 and>0.95 considered *satisfactory* and *excellent* respectively [58]. The Chi-square difference between the model and the data is routinely reported and should be small and non-significant, but is sample size dependent. It should, however, decrease in better fitting models. Since confirmatory factor analysis is intended to be theory-driven rather than exploratory, we proposed to test the four domain model on which the Individual Recovery Outcomes Counter is predicated, and the two factor model, or any close approximation of it revealed in the current study, reported in a previous study [23]. Confirmatory factor analysis was conducted using SSI International LISREL (V.9.10).

Results

Tests of data quality, distribution, stability, scaling assumptions, reliability, and validity

Analysis revealed a significant amount of missing data. Of *N*=2,680 baseline assessments *n* = 937 (34.9%) had missing data. There was no over-representation of any item among the missing data and, since sample size was not problematic, cases with missing values were deleted listwise leaving a final sample of *N* = 1,743 (see Table 1 for sample characteristics). The most highly scored item, indicating least recovery-related need, was safety and comfort, and the least highly scored item, indicating most need, was social network. Scaling assumptions were verified (see Table 2). Scale scores spanned the measurement continuum; mean item scores were largely dissimilar: repeated measures ANOVA revealed that 57/66 [86.4%] possible pairwise comparisons differed significantly; this is deemed acceptable when the intent of the tool is to extend the range of measurement to cover a wide range of health states [59]. Internal consistency was good (Cronbach's α= 0.85). There was little evidence of floor- or ceiling- effects with very few participants reporting either a maximum or minimum scale score. Data for all items, with the single exception of social network, which had a marked positive skew, were normally distributed. Transformation using Log 10 and Square Root methods did not resolve the issue. Mean *r* for test-retest reliability was .73 (*SD*=.06, range .61 to .82); six items fell within the good (*r*=.60 to .74 range) and six in the excellent (*r*>.75) range (*p*<.001 for all intraclass correlation coefficients). For the total score, *r*=.90 (95% CI .84, .94, *p*<.001).

>>Insert Table 2 about here<<

Rasch measurement testing of the Individual Recovery Outcomes Counter

Fit. At this stage all items fit the model with an overall standardized mean square item fit of .00 (*SD*=1.01), where values of 0 and 1 are expected. However, visual inspection of probability category curves and threshold scores revealed that all 12 items had a similar problem with fit. While the ordinal numbering of the response categories (i.e., 1 through 6)

were congruent with their imputed meaning (i.e., stronger endorsement of an item was associated with a higher total score and lower endorsement with the probability of a lower total score), Figure 3(A) shows that a response score of 4 was at no point the most probable outcome. In addition, for all items except social network, the Andrich threshold values for a response of 4 was greater than that for a response of 5. This suggested that a rating of 4 adds little of meaning to the item response categories, that they should be discarded, and items re-scored. However, a number of re-scoring options were possible; therefore, we calculated fit statistics and inspected probability curves for a range of alternative models (Figure 3B-F) to determine which, if any, scenario provided the most meaningful information and improved the quality of measurement taking place with these data. In detail, Figure 3(A) contains six hills, each indicating an original response option of 1-6 (1-2-3-4-5-6). Figure 3(B) illustrates five hills representing five category scoring; Figure 3(C) has four hills representing scale categories (1[234]56); Figure 3(D) shows responses to a tripartite model ([12][34][56]); Figure 3(E) illustrates four responses in the categories (1[23][45]6). Close inspection revealed that the four response model in Figure 3(D) depicts response categories that are ordered and working as intended. Inspection of infit and outfit statistics for the alternate solutions revealed that the 4-category model (Figure 3E) had the least underfit or overfit overall. Given that person (>2) and item (>3) separation; and person (>.8) and item (>.9) reliability were acceptable in all variations we therefore decided to re-score using the 4-category model depicted in Figure 3(E). Figure 4 shows the ordering of the threshold categories for each item using the 4-category model. The item map shows a person's expected score for each item as a function of the measure of personal recovery. The *x*-axis represents the theoretical continuum of the latent construct (less to more personal recovery) and the *y*-axis lists the items included. In this case, the six item response categories were collapsed to four response categories that are ordered and working as intended, thus further supporting the

case for re-scoring. As a result we have conducted all further analyses of the data based on the re-scored data. Normality of distribution was re-examined following re-scoring; data for all variables, including social network, now fell within acceptable limits.

With one exception, all scale items in the Individual Recovery Outcomes Counter fell within the reasonable range for infit and outfit (mean square =0.6-1.4). Social network was marginally outside this range (1.44 and 1.41 respectively) but not to a sufficient amount to degrade the scale and were therefore retained. Graphs showing the intraclass correlation coefficients were created for all items; observed values were located close to expected values with no marked deviation, and within, or at least very close to, 95% confidence intervals.

>>Insert Figure 2 about here<<

>>Insert Figure 3 about here<<

Targeting. Figure 4 shows the targeting of the sample to the 12 items and reveals that they capture just over three quarters (75.8%) of the sample. In particular, items did not adequately capture the people who report the lowest levels of personal recovery.

>>Insert Figure 4 about here<<

Dependency. High residual correlations (>.7) [60.61) may indicate local item dependency between pairs of items or persons. Residual correlations approached but did not exceed .7 (*Mdn* = .59, range .47 - .66) suggesting that up to half of random variance between items is shared.

Multidimensionality. Principal components analysis of the residuals revealed an Eigenvalue of 1.65 at the first contrast suggesting that the 12-item scale is unidimensional.

Item invariance. The DIF contrasts, comparing male and female responses ranged from -0.19 - 0.23, all well below the cut-off of 0.64 logits. Table 3 shows infit and outfit statistics. Higher patient measures were associated with higher item ratings; all point measure

correlations were in excess of .5 with the exception of social network (.46) which was sufficiently close to the expected correlation (.55) to be non-problematic.

Reliability. Possible scores on the item reliability index lie between 0 and 1 (>0.8 = *strongly acceptable*; [62]). Table 3 shows that item reliability on the 4-item scale used was 1.0. This indicates that items are adequately separating this sample along the measurement continuum.

>>Insert Table 3 about here<<

Factor structure of the Individual Recovery Outcomes Counter

Principal components analysis. Kaiser-Mayer-Olkin measure of sampling adequacy score was .907 indicating excellent adequacy of the data for factor analysis. The internal reliability for the 12-item scale (M = 35.6, SD = 7.99, α=.841) was good; in a single factor 'recovery' model 11 items loaded onto that factor >.5 (range .44-.72) with only social network failing to reach this threshold. Communalities in the single factor model ranged from .30 (social network) to .60 (exercise and activity); total communality was 5.60 and the percentage of variation explained by the model was .38 (38%). The best fitting model to emerge from principal components analysis was very similar to that previously extracted from a smaller sample [23] comprising two factors, one of eight and one of four items and accounted for 46.63% of total variance. There was no cross loading of factor items following rotation. Internal reliability of factor 1 (α=.809) and factor 2 (α=.636) was good and questionable respectively; application of the Spearman-Brown prophecy formula did not result in improved reliability. A 3-factor solution explained 55.89% of variance but there was significant cross-loading on a number of items. Four and five factor solutions both produced some factors with <3 items, and cross-loading items. Internal reliability of the four factors hypothesised by the tool's designers after application of the Spearman-Brown prophecy

formula was good (Home α=.835; Opportunity α= .862; Empowerment α=.895); and acceptable (People α= .798).

Confirmatory factor analysis. Calculation of the Mahalanobis distance revealed that assumption of multivariate normality was not supported because data for 30 (1.7%) individuals could be considered outliers. Subsequent inspection of the Chi-square-Mahalanobis scatter plot and distribution of univariate item data indicated that extreme scores were not due to an anomalous score on one variable. Inspection also indicated that none of the outliers were *error* outliers thus there is a case for removal of outlier data prior to confirmatory factor analysis since failure to meet the assumptions of multivariate normality can lead to overestimation of the Chi-square statistic and thus to increased chance of a Type I error [63]. Following guidelines [64] we checked whether removal of cases changed the resulting model(s) and report findings for the data both with and without removal of outliers. Further, we conducted confirmatory factor analysis on both models using the re-scored 4-category data and the original 6-category data to examine whether model fit worsened under the re-scored data condition. Inspection of 90% confidence intervals for RMSEAs revealed that no model was a significantly better fit than any other. However, under the two different scoring systems, the re-scored 4-category data always provides a better model fit than its equivalent scored on the original 6-category responses. Further, the two factor model revealed in the current study always provides a marginally better fitting model than the notional four domain model. This ultimately culminates in the re-scored 2-factor model which achieves an RMSEA of 0.051 (0.045-0.0565) verging on the margin of the threshold of an excellent fit (RMSEA=0.05). However, four domain model envisaged by the tool's developers was only marginally a less good fit (RMSEA=0.0536, 90%; CI 0.0477-0.0597). The very high correlation (.79) between the factors in the optimum model suggests that they may in fact comprise part of a super-ordinate personal recovery factor and not unrelated

latent variables. We tested this hypothesis by re-running the confirmatory factor analysis with covariance between the two factors set to zero. This resulted in a poorer fitting model (RMSEA=0.069) lending support to the idea that the items are all related to a single underlying construct.

Discussion

The current study has examined, in detail and with a considerably larger sample than previously conducted, the properties of the Individual Recovery Outcomes Counter, an outcomes and key-working tool for working with people with mental health problems towards their personal recovery. Results can inform the future development of the tool, and selection of the most appropriate model for use in clinical practice and outcomes reporting. The first important finding, revealed by Rasch analysis, is that the current 6-category scoring structure is problematic in the same way for each of the tool's 12 constituent items. A self-reported score of 4, representing a response of 'often', was at no point the most likely response for any item, thus rendering it redundant. Consideration of the reasons for this are warranted since one implication may be that future versions of the tool should use an amended scoring structure if it is the original scoring structure itself which is problematic. Analysis of alternative re-scoring methods revealed the best fitting solution involved collapsing categories 2 and 3, and 4 and 5, resulting in a 4-category structure: 1= *Never*; 2 = *Almost never/Sometimes*; 3= *Often/Most of the time*; 4 = *All of the time*. The solution itself suggests that the problem lies in the language of the original 6-category structure; while *never* and *all of the time* are clear and discrete categories, the others are less so and not obviously ordered. More importantly, it is not immediately obvious why the personal recovery construct should be rated solely along a continuum of frequency. While it is in keeping with the underlying recovery philosophy to concentrate on positive rather than negative aspects, e.g., by measuring in terms of possession of positive attributes or strengths rather than on

limitations, this should also encompass the intensity of the experience [65]. In other words, while one might 'often' experience good mental health and wellbeing, it does not follow that the relatively infrequent experience of less good mental health is not deeply distressing if the episode is characterised by great intensity. Hence, while re-scoring items in the current study appears to have improved model fit for a frequency-based model, the tool could potentially be improved further by development of a scoring structure which also requires consideration of personal recovery-related intensity. Nevertheless, in the current study the 4-category re-scoring produced a well-ordered tool with good fit statistics.

The second important finding is that the re-scored items measured only a portion of the personal recovery continuum. In particular, the Individual Recovery Outcomes Counter was unable to adequately target around a fifth of the sample who had total scores below item thresholds. More positively, the tool's items do successfully target around 75% of the relevant service user population and, therefore, has the potential to successfully track change for those who are furthest on their recovery journey. Nevertheless, the implication is that, for individuals in the bottom 20%, their true level of recovery is not captured; this means that those with most recovery-related needs are unlikely to be identified through use of the tool and therefore cannot be targeted for more intensive interventions. Further, those well below threshold scores are unlikely to demonstrate progress along the scale compared with those just below those thresholds; this might prove demoralising both for service users and workers when apparent progress fails to be captured. The conclusion to be drawn is that individual items may need to be amended in order to better target the full range of service users for whom it intends to have relevance. At this point, a note of caution is warranted since an amended scoring category-structure system that allows expression of recovery-intensity might, in itself, solve this problem and, hence, we suggest it will be beneficial to progress any tool redevelopment incrementally. Considerations to be made in adjusting, deleting or adding

items should include whether specific types of item will improve targeting, and whether there is item redundancy. In respect of the former, priority should be given to development of items that are sensitive to reduced levels of personal recovery.

While we conducted exploratory and confirmatory factor analysis ostensibly to test competing theories about the factor structure of the Individual Recovery Outcomes Counter , we found only partial support for either. Further, it might be considered that even partial support contradicts indications from the Rasch analysis that the tool is unidimensional. While a 2-factor intrapersonal/interpersonal structure was supported to an extent, the component factors were highly correlated suggesting that both represent a single, super-ordinate factor. Further, while the structure was very similar to that revealed in Monger et al.'s [23] analysis it was not identical. The relocation of the hope for the future item from what had been interpreted as an interpersonal, outward/forward looking factor to an intrapersonal self-reflection/change factor and that of the physical health item in the opposite direction is worthy of consideration. The former change, in particular, brings into question the 'forward looking' aspect of this interpretation; indeed, no obvious intuitive solution has occurred to us and this in our view further strengthens the case for a unidimensional scale.

We suggest that a pragmatic rather than prescriptive approach is warranted; essentially, rather than make categorical statements about the tool as definitively unidimensional or multidimensional, findings should be used as a diagnostic aide and guide for its future development. Currently, the weight of evidence supports the unidimensionality of the tool. Nevertheless, the notional four domain structure is user-friendly and provides an intuitively appealing approach which provides an aide-memoire for both workers and service users. However, statistically it does not provide the best explanation of the data. The 2-factor structure, similarly, does not provide the best empirical solution but does suggest future routes of development. Given that development of the tool has not included statistical testing

of a larger pool of potential constituent items it is possible that the emergence of a 2-factor structure in the principal components analysis might be strengthened by the addition of more items similar to those in factor 2. However, while not statistically significant, those items found to constitute the previously-titled interpersonal factor in principal components analsysis are definitely positioned to the left of the personal recovery continuum (the fifth, seventh, ninth and twelfth most 'difficult' items) relative to the previously-titled intrapersonal items. We can conclude, then, that inclusion of more of this type of item is unlikely to extend the targeting range of the tool. We also note at this juncture that two pairs of items (valuing myself and mental health, and life skills and personal network) sit at the same part of the personal recovery continuum as one another and there may be some redundancy.

Since the Individual Recovery Outcomes Counter is intended to inform the therapeutic key-worker - service user dialogue as well as functioning as an outcome tool, any added or amended items will need to be of true clinical as well as of statistical value; conversely, removal of items should be done cautiously where they retain clinical value. Simply adding items to the tool for statistical expedience may not add practical utility and should be avoided since this would simply increase item redundancy [66]. A clear corollary of this is that further generation of ideas for new items should be led by experts by experience once they have been apprised of the findings of the current study. Nevertheless, examination of the wider mental health recovery outcomes literature suggests some potentially fruitful issues to consider for inclusion might be an item related to work, since the current item purpose and direction may not sufficiently capture the unique value to potentially be made by gainful paid employment. A potential barrier is the relatively small proportion (12%) of service users engaged in paid employment which might exert a ceiling effect on responses. It is also necessary to consider that, while many people with mental health conditions desire to be employed, it may be counterproductive to set expectations that are perceived to exert

pressure to take unsuitable employment. Candidate variables might also include items such as 'Giving it back' [67]. Developers should aim not to make any new version unwieldy and, given the status of the tool in shaping the therapeutic conversation, should first and foremost be guided by what is helpful to the patient and the relationship. We suggest that evidence for the notional four domain structure is sufficiently strong for it to be retained in key-working materials and outcomes tools. However, it may be more useful from a development perspective to consider the tool as comprising two recovery-related strands that requires some balancing in favor of intrapersonal-type items. The challenge, of course, will be in integrating these approaches when these and scoring-related modifications are made. Further examination of factor structure will be warranted following any modifications made to the Individual Recovery Outcomes Counter.

Given the changes required, it may seem moot at present to address whether the current findings on factor structure can inform clinical use of the tool. For example, could either the two factor intrapersonal/interpersonal or scores derived from the notional four domain structure be used to inform interventions or to be meaningfully reported in outcomes data? Clearly this would be an aim of further development but at present it is not possible to recommend this approach until issues of targeting, scoring, and potential multidimensionality have been resolved. Finally, since the scale-items accounted for less than half (47%) of the variance in the sample scores, it should be fruitful to listen carefully to what users say about other issues that affect their personal sense of recovery and help them with these issues where possible. Recording of additional issues might aid identification of potential new items.

Limitations

The study depended upon routinely collected assessment data gathered in the day-to-day work milieu. There was a considerable number of incomplete assessments; it may be unavoidable in the context of routine assessment with people with mental health problems

that a considerable proportion do not initially engage with services. Nevertheless, we could detect no systematic reason for non-engagement in terms of demographic variables. The data collected here was, in fact, subject to some quality control in terms of training. A further potential limitation is the patchiness of some of the descriptive demographic and clinical data whose inclusion would have allowed us to better describe the sample and to test DIF for variables other than gender: it is possible that responses may have differed between, for example, people with psychotic disorders and those with anxiety and non-psychotic depressive episodes. Future work should aim to gather more complete demographic and clinical data to better address this question.

Conclusion

The current study provides further support for the use of the Individual Recovery Outcomes Counter as a unidimensional measurement of recovery; this is qualified support, however, and further development is required in order for the tool to more adequately capture the recovery of those who are at most disadvantage. This paper sets out a programme of work to achieve this.

Acknowledgments: The authors would like to thank Nigel Henderson and Jane Cumming (Penumbra)

Declaration of interest: None

References

[1] Slade M. Policy rationale. In: M. Slade. Personal Recovery and Mental Illness (p.74-77). Cambridge (UK): Cambridge University Press; 2009.

[2] Tomes N. The patient as a policy factor: A historical case study of the consumer/survivor movement in mental health. Health Aff. 2006;25:720-729. doi: 10.1377/hlthaff.25.3.720

[3] Deegan PE. Recovery and empowerment for people with psychiatric disabilities. Soc Work Health Care. 1997;25:11-24. doi:10.1300/J010v25n03_02

[4] Jacobson N, Greenley D. What Is Recovery? A conceptual model and explication. Psychiatr Serv. 2001;52:482-485. doi: 10.1176/appi.ps.52.4.482

[5] Topor A, Borg M, Di Girolamo S, et al. Not just an individual journey: social aspects of recovery. Int J Soc Psychiatry.2011;57:90-99. doi: 10.1177/0020764009345062

[6] Davidson L, O'Connell MJ, Tondora J, et al. (2005). Recovery in serious mental illness: a new wine or just a new bottle? Prof Psychol Res Pract. 2005;36:480-487. doi:10.1037/0735-7028.36.5.480

[7] Stickley T, Wright N. The British research evidence for recovery, papers published between 2006 and 2009 (inclusive). Part One: a review of the peer-reviewed literature using a systematic approach. J Psychiatr Ment Health Nurs. 2011a;18:247-256. doi:10.1111/j.1365-2850.2010.01662.x

[8] Stickley T, Wright N. The British research evidence for recovery, papers published between 2006 and 2009 (inclusive). Part Two: a review of the grey literature including book chapters and policy documents. J Psychiatr Ment Health Nurs. 2011b;18: 297-307. doi: 10.1111/j.1365-2850.2010.01663.x

[9] Anthony WA. Recovery from mental illness: the guiding vision of the mental health service system in the 1990s. Psychosoc Rehabil J. 1993;16:11-23. doi: http://dx.doi.org/10.1037/h0095655

[10] Carson J, McManus G, Chander A. Recovery: a selective review of the literature and resources. Ment Health Soc Incl. 2010;14:35-44. doi: http://dx.doi.org/10.5042/mhsi.2010.0068

[11] Ion R, Monger B, Hardie S, et al. A tool to measure progress and outcome in recovery. *Br J Ment Health Nurs. 2013;2*:211-215. doi:10.12968/bjmh.2013.2.4.211

[12] Le Boutillier C, Chevalier A, Lawrence V, et al. (2015). Staff understanding of recovery-orientated mental health practice: A systematic review and narrative synthesis. Implement Sci. 2015;10:87-87. doi:10.1186/s13012-015-0275-4

[13] Salzmann-Erikson M. (2013). An integrative review of what contributes to personal recovery in psychiatric disabilities. *Issues in Mental Health Nursing, 34,* 185-191. doi:10.3109/01612840.2012.737892

[14] Shepherd A, Doyle M, Sanders C, et al. Personal recovery within forensic settings: Systematic review and meta-synthesis of qualitative methods studies. Crim Behav Ment Health. 2016;26:59-75. doi:10.1002/cbm.1966

[15] Van Lith T, Schofield MJ, Fenner P. Identifying the evidence-base for art-based practices and their potential benefit for mental health recovery: A critical review. Disabil Rehabil. 2013;35:1309-1323. doi:10.3109/09638288.2012.732188.

[16] Scheyett A, DeLuca J, Morgan C. Recovery in severe mental illnesses: A literature review of recovery measures. Soc Work Res. 2013;37:286-303. doi:10.1093/swr/svt018

[17] Shanks V, Williams J, Leamy M, et al. Measures of personal recovery: A systematic review. Psychiatr Serv. 2013;64:974-980. doi:10.1176/appi.ps.005012012

[18] Monger B, Hardie SM, Ion R, et al. The Individual Recovery Outcomes Counter: Preliminary validation of a personal recovery measure. Psychiatri.2013;37:221-227. doi:10.1192/pb.bp.112.041889

[19] Tractenberg RE. Classical and modern measurement theories, patient reports, and clinical outcomes. Contemp Clin Trials. 2011;31:1-3. doi: 10.1016/S1551-7144(09)00212-2

[20] Urbina S. Essentials of psychological testing. Hoboken (NJ): John Wiley & Sons Inc; 2004.

[21] Cano SJ, Hobart JC. The problem with health measurement. Patient Preference Adherence. 2011;5:279 –290. doi: http://dx.doi.org/10.2147/PPA.S14399

[22] Cano SJ, Mayhew A, Glanzman AM, et al. Rasch analysis of clinical outcome measures in spinal muscular atrophy. Muscle Nerve. 2014;49:422–430. doi: http://dx.doi.org/10.1002/mus.23937

[23] Monger B, Hardie SM, Ion R, et al. The Individual Recovery Outcomes Counter: Preliminary validation of a personal recovery measure. Psychiatri. 2013;37:221-227. doi:10.1192/pb.bp.112.041889

[24] Christensen KB, Engelhard Jr, J. Salzberger, T. Ask the experts: Rasch vs. factor analysis. Rasch Measur Trans. 2012;26:1373-1378.

[25] Waugh RF, Chapman ES. An analysis of dimensionality using factor analysis (true-score theory) and rasch measurement: What is the difference? Which method is better? J App Measur.2005;6:80–99.

[26] Yu CH, Popp SO, Digangi S, Jannasch-Pennell A. Assessing unidimensionality : a comparison of Rasch Modeling,Parallel Analysis, and TETRAD. Pract Assessment Res Eval. 2007;12:1–19.

[27] Comrey AL, Lee HB. A first course in factor analysis. Hillsdale (NJ): Erlbaum; 1992.

[28] Hair JFJ, Anderson RE, Tatham RL, et al. Multivariate data analysis. Cambridge (UK): Pearson; 2010.

[29] Giffort D, Schmook A, Woody C. Recovery Assessment Scale. Chicago (IL): Department of Mental Health; 1995.

[30] Corrigan PW, Salzer M, Ralph RO, et al. Examining the factor structure of the Recovery Assessment Scale. Schizophr Bull. 2004;30:1035–1041. doi:10.1093/oxfordjournals.schbul.a007118

[31] Sederer LI, Dickey B, Eisen SV. Behavior and Symptom Identification Scale (BASIS-32). In: Ll. Sederer & B. Dickey (Eds). Outcomes Assessment in Clinical Practice (p.65–69). Baltimore,(MD): Williams & Wilkins; 1996.

[32] Eisen SV, Wilcox M, Leff HS. Assessing behavioural health outcomes in outpatient programs: reliability and validity of the BASIS-32. J Behav Health Serv Res. 1999;26:5-17. doi: 10.1007/BF02287790

[33] Bond TG, Fox CM. Applying the Rasch model. fundamental measurement in the human sciences (2nd Edition). New York (NY): Routledge; 2007.

[34] Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol Assess.1995;6:284–290. doi:10.1037/1040-3590.6.4.284 doi:10.1093/oxfordjournals.schbul.a007118

[35] Rasch G. Probabilistic models for some intelligence and attainment. Chicago (IL): University of Chicago Press; 1980.

[36] Massof DW. Is the partial credit model a Rasch model? J Appl Measur. 2012;13:114-131.

[37] Linacre JM. Partial Credit Models" (PCM) and "Rating Scale Models" (RSM). Rasch Measur Trans. 2000;14:768.

[38] Barbic SP, Bartlett SJ, Mayo NE. Emotional vitality in caregivers: application of Rasch Measurement Theory with secondary data to development and test a new measure. Clin Rehabil. 2015;29:705-716. doi:10.1177/0269215514552503

[39] Barbic SP, Kidd SA, Davidson L, et al. Validation of the brief version of the Recovery Self-Assessment (RSA-B) using Rasch measurement theory. Psychiatr Rehabil J. 2015 38:349-358. doi: 10.1037/prj0000139

[40] Covic T, Pallant J, Conaghan P, et al. A longitudinal evaluation of the Center for Epidemiologic Studies-Depression scale (CES-D) in a rheumatoid arthritis population using Rasch analysis. Health Qual Life Outcomes. 2007;5:41. doi: 10.1186/1477-7525-5-41

[41] Pallant JF, Tennant A. An introduction to the Rasch measurement model: An example using the Hospital Anxiety and Depression Scale (HADS). Br J Clin Psychol. 2007;46:1–18. doi: 10.1348/014466506X96931

[42] Linacre JM. Understanding Rasch measurement: Optimizing category effectiveness. J Appl Measur. 2002;3:85-106 .

[43] Wright BD, Linacre JM, Gustafson J, et al. Reasonable mean-square fit values. Rasch Measur Trans. 1994; 8:370.

[44] Linacre JM. Winsteps Rasch tutorial 2 [Internet]. 2012 . [cited 2015 Jan 5]. Available from: http://www.winsteps.com/a/winsteps-tutorial-2.pdf

[45] Wright BD, Masters GN. Rating scale analysis. Chicago (IL): MESA Press; 1982.

[46] Sick J. Rasch measurement in language education part 6: Rasch measurement and factor analysis. JALT Test Evaluation SIG Newsl. 2011;15:15-17.

[47] Linacre JM. Structure in Rasch residuals: why principal components analysis (PCA)? Rasch Measur Trans. 1998; 12:636.

[48] Linacre JM. Data variance explained by Rasch measures. Rasch Measur Trans. 2006;20:1045.

[49] Smith Jr. EV. Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. J Appl Measur. 2002;3:205-231.

[50]Linacre JM. DIF - DPF - bias - interactions concepts [Internet]. no date. [cited 2015 Jan 15]. Available from: http://www.winsteps.com/winman/difconcepts.htm

[51] Andrich D. An index of person separation in latent trait theory, the traditional KR. 20 index, and the Guttman scale response pattern. Educ Res Perspect. 1982;9:95-104.

[52] Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychom. 1951;16:297-334.

[53] Stewart DW. The application and misapplication of factor analysis in marketing research. J Mark Res, 1981;18:51-62.

[54] Costello A, Osborne J. Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. Pract Assess Res Evaluation, 2005;10:173-178.

[55] Tabachnick BG, Fidell LS. Using multivariate statistics. Upper Saddle River (NJ): Pearson, Allyn, and Bacon; 2007.

[56] George D, Mallery P. SPSS for Windows step by step: A simple guide and reference. 11.0 update (4th ed.). Boston (MA): Allyn & Bacon; 2003.

[57] Wells C, Wollack J. An instructor's guide to understanding test reliability [Internet]. 2003. [cited 2016 December 2]. Available from: http://testing.wisc.edu/Reliability.pdf

[58] Byrne B. Structural Equation Modeling with AMOS. New York (NY): Routledge; 2010.

[59] Ware JE, Gandek B. (1998). Methods for testing data quality, scaling assumptions, and reliability: The IQOLA project approach. J Clin Epidemiol. *1998;51:*945-952. doi: http://dx.doi.org/10.1016/S0895-4356(98)00085-7

[60] Yen,WM. Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. Appl Psychol Measur. 1984;8:125-145. doi: 10.1177/014662168400800201

[61] Yen WM. Scaling performance assessments: Strategies for managing local item dependence. J Educ Measur. 1993;30:187-213. doi: 10.1111/j.1745-3984.1993.tb00423.x

[62] Fox CM, Jones JA. Uses of Rasch modeling in counseling psychology research. J Couns Psychol. 1998;45:30-45. doi: http://dx.doi.org/10.1037/0022-0167.45.1.30

[63] Curran PJ, West SG, Finch JF. The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. Psychol Methods. 1996;1:16-29. doi: http://dx.doi.org/10.1037/1082-989X.1.1.16

[64] Aguinis H, Gottfredson RK, Joo H. Best-practice recommendations for defining, identifying, and handling outliers. Organ Res Methods. 2013;16:270-301. doi: 10.1177/1094428112470848

[65] Dela Cruz AM, Bernstein IH, Greer TL, et al. Self-rated measure of pain frequency, intensity, and burden: psychometric properties of a new instrument for the assessment of pain. J Psychiatr Res.2014;59:155–160. doi: http://doi.org/10.1016/j.jpsychires.2014.08.003

[66] Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. New York (NY): Oxford University Press; 1989.

[67] Ridgway P. Re-storying psychiatric disability: learning from first person recovery narratives. Psychiatr Rehabil J. 2001;24:335-343. doi: http://dx.doi.org/10.1037/h0095071

Figure captions

Figure 1: Individual Recovery Outcomes Counter notional four domain (HOPE) model with example scoring and item descriptors

Figure 2: Service user-Item Threshold Distribution

Accompanying text:

Distribution of I.ROC items obtained by converting raw scores into logits. The *x*-axis represents the level of personal recovery continuum from low to high. The top bars (above the *x*-axis) represent the distribution of people in the sample, whereas the bottom bars (below the *x*-axis) represent items. Ideal targeting would depict a range of item and people covering the whole breadth of the scale. The figure shows that there are few items (bottom bars) that are covering the people (top bars) at the low end of the continuum (those lowest on the continuum of personal recovery). The scale is well-targeted at those scoring between -1.2 and +1.8 logits. The measurement gaps in the personal recovery continuum are shown by the 2-way block arrows. Those scoring above the measurement threshold represent 2.9% of the total sample while those scoring below represent 21.2% of the sample. Therefore, a solution is required to capture the lower end of the personal recovery construct for this sample. In addition, several items are capturing the level of personal recovery of the same subgroup of patients.

Key: [1]. Social Network; [2]. Valuing myself; [3]. Mental health [4]. Purpose & direction; [5]. Hope for the future; [6]. Physical health;[7].Self management; [8]. Exercise & activity; [9]. Participation & control; [10]. Personal network;; [11]. Life skills; [12]. Safety & Comfort.

Figure 3: Probability category curves of item 1 (Mental Health)

Accompanying text:

In all examples the responses are correctly ordered, but in the original scoring model (A) a score of 4 is at no time most probable. Models B to F represent alternative re-scoring methods. Collapsing scores of 4 and 5 (Model B), 3,4, and 5 (C), 2 and 3 and 3 and 4 (D) and 1 and 2, 3 and 4, and 5 and 6 (E) all lead to ordered solutions. Visual inspection suggests Model D to be the most satisfactory and is supported by inspection of fit statistics (See Table 3).

Figure 4: Item map

Accompanying text:

Item map showing an individual's expected score to each item as a function of the measure of personal recovery. The *x*-axis represents the theoretical continuum of the latent construct (less to more personal recovery) measured in logits. The *y*-axis lists the I.ROC items in terms of more (Safety and comfort) to less (Social network) personal recovery. In this case the 6-item response categories were collapsed into 4 (i.e., collapse responses 2 & 3 and 4 & 5 into one category each (2 and 3) and transform response 6 to 4). The figure depicts response

categories that are ordered and working as intended, suggesting a 4 item category response may be more favourable for this sample.

Table 1

Participant Characteristics

		n	(%)
Gender	Male	787	(45.1)
	Female	957	(54.9)
Ethnicity	White British	927	(53.2)
	Asian/ Asian mixed/ Asian other	17	(1.0)
	White other	14	(0.8)
	African-Caribbean	4	(0.2)
	Not known/prefer not to say	782	(44.8)
Referral from	NHS	316	(18.1)
	Self/private	287	(16.5)
	Social work	273	(15.7)
	Housing	239	(13.7)
	Community/Independent service	180	(10.3)
	Education authority	137	(7.9)
	General Practitioner	81	(4.6)
	Self	1	(0.1)
	Other	200	(11.5)
	Not recorded	70	(4.0)
Reason for leaving service	Moved	548	(31.4)
	Did not engage	433	(24.8)
	Still in service	412	(23.6)
	No longer meeting client needs	213	(12.2)
	Disengaged with service	134	(7.7)
	Died	4	(0.2)
Employment status	Unemployed	1138	(65.3)
	Student	340	(19.5)
	Employed	209	(12.0)
	Other	57	(3.3)
Length in service (Days)	*M*	240.3	
	SD	290.1	
	Range	1-3416	
Age at assessment (Years)	*M*	37.9	
	SD	14.9	
	Range	18-81	

Table 2

Analysis of data quality, scaling assumptions, targeting, and reliability

Psychometric property	Total
Data quality:	
Missing data (%)	34.9
Computable scale scores	1743
Scale assumptions:	
Item scores: *M* (range)	3.67 (2.17-4.09)
Item *SD* range	1.16-1.55
Targeting:	
Mean score (*SD*)	38.24 (10.17)
Possible score range[a]	12-72
Observed score range	12-72
Floor/ ceiling effect[b]	<1/<1
Rating scale score:	Of 20,916 observations:
1 = *Never*	14%
2 = *Almost never*	20%
3 = *Sometimes*	29%
4 = *Often*	15%
5 = *Most of the time*	14%
6 = *All the time*	8%
Reliability:	
Cronbach's alpha	.85
M (*SD*, Range) inter-item correlation byitem	.32 (.07, .14-.43)
Item-total correlation *M* (*SD*, Range)	.61 (.08, ,47 - .70)

[a] Higher scores represent higher personal recovery [b] Floor effect = % receiving a score of 12 (lowest personal recovery); ceiling effect = % receiving a score of 72 (highest possible personal recovery orientation total score on the original Individual Recovery Outcomes Counter 6-point scale)

Table 3

Item fit statistics

Item	Total Score	Total Count	Measure	SE	Infit MNSQ[a]	Infit ZSTD[b]	Outfit MNSQ[c]	Outfit ZSTD[d]
Mental Health	3,473	1,743	.39	.04	.67	-9.9	.68	-9.9
Life Skills	4,461	1,743	-.66	.04	.82	-6.1	.81	-6.2
Safety & Comfort	4,857	1,743	-1.21	.04	1.30	8.7	1.31	8.8
Physical Health	4,000	1,743	.00	.04	.92	-2.5	.93	-2.1
Exercise & Activity	4,072	1,743	-.10	.04	1.27	7.5	1.27	7.4
Purpose & Direction	3,755	1,743	.37	.04	.99	-.3	.99	-.2
Personal Network	4,437	1,743	-.62	.04	1.23	6.7	1.24	6.9
Social Network	3,043	1,743	1.55	.04	1.46	9.9	1.41	9.9
Valuing Myself	3,548	1,743	.70	.04	.88	-3.5	.88	-3.8
Participation & Control	4,424	1,743	-.61	.04	.95	-1.6	.96	-1.4
Self-management	4,020	1,743	-.03	.04	.67	-9.9	.67	-9.9
Hope For The Future	3,857	1,743	.22	.04	.86	-4.3	.87	-4.2

[a] Infit MNSQ (mean square) = outlier sensitive fit statistic; it is sensitive to unexpected observations by service users on items that are relatively very hard or very easy for them (and vice versa); [c] Outfit MNSQ = Inlier-pattern-sensitive fit statistic, sensitive to unexpected responses by service users on items that are roughly targeted on them (and vice versa). Infit and Outfit ZSTD[b,d] (standardised Z-score or 't-statistic') report statistical significance (1.96 = $p<.05$) of MNSQ values occurring by chance when the data fit the Rasch model. Infit and Outfit MNSQ values close to 1.0 indicate acceptable fit and that items are productive for measurement (1.5-2.0 unproductive for measurement and >2.0 are degrading to the scale measurement). Significant t-statistics can be ignored if MNSQ is acceptable [60]