Title: An initial evaluation of the Global Review Form as an approach to measuring individual change

Authors: Natalie Bentley[a,b], Jason Davies[a,b,c], Joselyn Sellen[b] & Richard Maggs[a]

[a] Abertawe Bro Morgannwg University Health Board, Taith Newydd Low Secure Service, Glanrhyd Hospital, Bridgend. UK
[b] Department of Applied Psychology, Cardiff Metropolitan University, Cardiff, UK
[c] Department of Psychology, Swansea University, Swansea, UK

Corresponding Author

Jason Davies

Department of Psychology, College of Human and Health Sciences, Swansea University, Singleton Park, Swansea. SA2 8PP

Email: Jason.davies@swansea.ac..uk

Abstract

Background: Routine assessment of individual change in forensic mental health services is increasingly recognised as important. However, existing tools have been criticised and their periodic use make them unsuited to directly measure the impact of interventions. This paper describes the initial evaluation of the Global Review Form (GRF) as a framework for measuring change over time. Specifically, measurement properties, feasibility and usefulness in routine practice are examined.

Method: 28 male service users in three distinct areas of an adult secure service (low secure, locked rehabilitation and high relational support housing) were rated over a twenty week period by their multidisciplinary teams.

Findings: The GRF showed promising construct validity and appropriate stability and sensitivity to change across time. It enabled measurement and understanding of individual change over time. Staff feedback suggested the GRF is a useable and practical outcome measuring tool.

Conclusions: The GRF shows promise for use as a routine outcome monitoring tool within forensic mental health services.

An initial evaluation of the Global Review Form as an approach to measuring individual change

Measuring outcomes in mental health services is becoming increasingly important to both practitioners and policy makers (Yiend, Chambers, Burns, Doll, Fazel, Kaur, Sutton & Fitzpatric, 2011) with UK Government policies on mental health practice emphasising the importance of routinely measuring individual patient outcomes (Department of Health, 1999). Davies, Howells, and Jones (2007) highlight the importance of measuring outcomes within a forensic mental health setting. Reasons for this include that some interventions in forensic services can make people worse (Jones, 2007); the high cost of treatments and the limited availability of interventions within forensic services. However, outcome measurement in forensic mental health has been criticised. Thornely and Adams (1998) note the lack of appropriate outcome measures suitable for use in this setting whilst Cohen and Eastman (2000) report that outcome measurement in forensic mental health typically focuses on criminal justice outcomes (e.g. reconviction) with limited emphasis on change at the individual level.

Two instruments are widely used in forensic mental health services; the HCR-20v2 (Webster, Douglas, Eaves & Hart, 1997) to monitor factors in relation to risk and the HoNOS-secure (Wing, Beevor, Park, Haddon, Burns & Burns, 1998) to measure various facets of clinical and security need over time. Despite research showing such measures to have good reliability and validity (e.g., Wing et al., 1998), there are some limitations with these tools for routine change measurement in a forensic mental health service. Stein (1999), states that the HoNOS is not specific enough to input into specific care plans, and lacks sensitivity to smaller changes. In comparison, the HCR-20v3 (Douglas, Hart, Webster, & Belfrage, 2013) has not been designed as an outcome measure although it (and the earlier version) is often used to monitor changes in risk factors over time. Further, the HCR-20v3 focuses on those factors associated with a risk of violence, thus many areas of interest and functioning are not included. Other outcome measures which have been developed for use within a forensic mental health setting have also been criticised (Cure and Adams, 2000). In their recent review of outcome measures in forensic settings, Shinkfield & Ogloff (2014) identified six tools (out of a pool of 19) that met their evaluation criteria. Their goal was to identify tools which assess most or all of four outcome areas namely, functioning, recovery, risk and placement pathway. Along with the HONOS, the shortlisted six included the DUNDRUM toolkit (e.g. Davoren et al, 2015a, 2015b). This suite of tools provides a range of assessment (including patient self ratings of recovery) that has been shown to have some predictive utility in relation to future placement.

Studies evaluating the psychometric properties of outcome measures in mental health settings have identified several criteria which are fundamental in determining the efficacy of an outcome measure (e.g. Burgess, Pirkis, Coombs, & Rosen, 2010; Donnelly, Scott, McGilloway, O'Neill, Williams & Slade, 2011). Burgess et al, (2010) provide a series of benchmarks against which to assess recovery orientated tools (a similar approach was subsequently adopted by Shinkfield and Ogloff, 2014 in relation to forensic outcome measures). These include brevity (fewer than 50 items) and sound psychometric properties. Further, Donnelly et al., (2011) suggest that outcome measurement tools should demonstrate good validity, reliability and interpretability and be sensitive enough to capture change over time; this should also address the clinical meaningfulness of the individual change measured (Jacobson & Truax, 1991). To this end, Davies et al., (2007) propose that in evaluating individual change and its clinical relevance, single case methodologies can be useful. One feature of such approaches is frequently repeated measurement over time, something that forensic mental health tools do not readily lend themselves to (e.g. DUNDRUM – includes item ratings that need 12 months+ to achieve, HONOS-s – suggested re-rating period of 6 months or change in setting).

Slade, Thornicroft and Glover (1999) suggest that establishing the feasibility of an outcome measure is essential in order to ensure the instrument is "suitable for routine, sustained and meaningful use within a typical clinical setting". However, the widespread practice of relying on individual clinical judgement in relation to change remains prevalent. This is problematic; for example, Nicholson and Norwood (2000), highlight that the reasoning behind expert opinions within reports is often difficult to determine, and such views often rely on idiosyncratic clinical judgement, self-report, heuristic's and memory (Wettstein, 2005). Therefore it is essential to improve methods for quantifying, and correcting for biases in experts decision making process (Malsch & Freckelton, 2005).

The Global Review Form (GRF) was developed to enable frequent repeated measurement to quantify individual change over time amongst those within local forensic mental health services (Davies, 2011; Davies & Maggs, 2009). In this way it differs from the tools noted above (e.g. DUNDRUM, HONOS-s) as it is expected that ratings will be made every 2-4 weeks. Its purpose is to facilitate progress monitoring across eight domains, evidence the effectiveness of care and treatment, and aid clinical decision making within a low secure forensic mental health setting. Teams are encouraged to prioritise those domains where a score of -1 or below has been allocated whilst continuing with treatment and monitoring in other areas.   The GRF is grounded in the structured professional judgement approach to

clinical decision making (Kropp & Hart, 2000) which combines the strengths of unstructured clinical decision making and actuarial assessment (Douglas & Kropp 2002). The GRF is intended to aid clinical decision making by routinely collating the clinical teams' professional view of an individual's functioning across a range of domains such as mental health, treatment engagement and self care. An early version included scales relating to risk (to others, self, of relapse and absconscion). However, as a range of risk specific tools were also being used (e.g. HCR-20 for violence risk) the risk scales were removed with the expectation that clinical teams adopt risk assessment methods in addition to completing the GRF. Thus, the scales reflect three of the four domains of functioning, described by Shinkfield and Ogloff (2014) namely various aspects of functioning; placement pathway, and the service perspective of recovery. Work is currently underway to gather client perspectives of recovery using the GRF framework. The ratings are intended to inform care plans, back up opinions in Tribunals and other formalised scrutiny processes and provide a defensible means to validate clinical decisions. In addition, identifying the patterns of change within an individual may aid in understanding the effects of specific interventions or events – this could provide clinicians with a valuable insight into how and why change has occurred (Davies, Jones, & Howells, 2010). At a group level, data from the GRF might assist services to identify patterns in the typical profile of change, who the service benefits the most and the impact of environmental / contextual decisions and interventions.

The GRF is based on goal attainment scaling (GAS) principles (Kiresuk and Sherman, 1969) which have been shown to be a reliable and direct method for assessing the effectiveness of an intervention or service, and for measuring individual change (Smith, 1994). Following the principles of GAS, the GRF consists of eight domains each rated on a 7 point scale, ranging from +3 to -3, with 0 representing an adequate level of functioning. The GRF is intended to be used within multi-professional team review meetings where consensus ratings can be provided on each domain for an individual.

Aims and objectives

The aims of this study were to evaluate the measurement properties and feasibility of the GRF for routine use as a measure of change.

More specifically this research sought to

a) Evaluate the GRF's construct validity to determine whether the GRF effectively captures the current functioning of service users;

b) Evaluate the GRF sensitivity to change over time and its ability to capture clinically meaningful change;

c) Determine the interpretability and usefulness of the information obtained from the GRF in assessing individual change through a single case report;

d) Assess the feasibility of the GRF for use as a routine outcome measure via questionnaire feedback from the professionals using the tool.

Methodology

Ethical considerations

As the data was routinely collected information used by the clinical team, the respective NHS Research and Development Department classified this study as service evaluation. Ethical scrutiny was provided by a University Ethics Panel and this work was undertaken according to best practice standards e.g. British Psychological Society's Code of Ethics and Conduct (British Psychological Society, 2009).

Participants

Two sources of data were used in this study;

Routine individual monitoring

Twenty eight male inpatients from a male secure NHS mental health service were rated by their Multi-disciplinary team (MDT) over a maximum 20 week period using the GRF. The secure service consists of three clinical areas; a low secure ward, a locked rehabilitation ward and a step down facility. All provide rehabilitation to service users who are diagnosed with a serious mental disorder and require the provision of security (relational and / or physical) in relation to their risk. These services can be differentiated according to their physical, and relational security, i.e.: low secure e.g. anti-climb perimeter fence, airlock, daytime (nurse) staffing of 1 staff : 2.3 patients; locked rehabilitation e.g. standard 6 foot perimeter fencing, single locked door, 1 staff : 3.6 patients; and step down e.g. 3 foot garden wall, open door policy, 1 staff : 4 patients. All areas provide multi-professional care with the service provision within the low secure ward having been previously described (Davies, Maggs and Lewis, 2010).   Whilst those in the low secure ward typically stay for up to 18 months, those within the locked rehabilitation area have the longest inpatient stays, often longer than 2 years. The step down facility offers a graded transition from the inpatient settings to community living with the typical stay less than 12 months in duration. The multi-professional teams making the ratings typically included a consultant psychiatrist, a senior

nurse, a clinical or forensic psychologist and an occupational therapist. The same team rated those in low secure and step down provision with a second team rating those in locked rehabilitation.

As per inclusion criteria for these services, all 28 participants were male, aged between 18 and 65 and all had a diagnosis of serious mental illness (schizophrenia or schizoaffective disorder). None had a formal co-morbid diagnosis of personality disorder although five had significant problematic personality traits. The vast majority (25) had a history of problematic alcohol use and / or a substance misuse history. The majority of service users were detained under the Mental Health Act (2007), under civil sections (s3) or criminal justice sections (e.g. s37); two service users were informal patients. Many of the individuals had been in contact with the criminal justice system at some point in the past. All service users were expected to remain within the service during the data collection period.

GRF staff feedback

An opportunity sample of 14 qualified mental health professionals working within the service, and who had been involved in using the tool were approached to provide feedback on the GRF. Ten completed questionnaires were returned for analysis.

Design

In order to adequately address the research aims, and to provide a thorough analysis of the GRF, this study used a mixed methods design (Tashakkori & Teddlie, 1998). A quantitative design was used to examine some of the measurement properties of the tool whilst a descriptive and qualitative approach was used to examine feedback on the tool's feasibility and usefulness. Finally, a single case design was adopted to further examine the idiographic utility of the tool.

Materials

The Global Review Form

The Global Review Form (Davies, 2011; Davies & Maggs, 2009), is designed to measure individual change over time across 8 different domains, namely: medication; symptom management; general engagement; leave; relationships and support; substance misuse;

occupations and environmental input. Each domain is rated, by the clinical team, on a 7 point scale (+3 to -3) with anchor descriptions provided for each point on the scale. Ratings at +3 represent the greatest level of functioning with -3 representing the lowest functioning that can be recorded. The GRF also has a 'traffic light' colour coded system; with amber (+1, 0, -1) signifying adequate functioning, green (+2, +3) signifying above expected level of functioning (which may prompt discharge discussions), and red (-2, -3) highlighting an individual is below expected functioning (and may need review or additional resources or input). A rating sheet is used to record the consensus ratings produced by the team. Copies of the user manual and rating forms can be obtained from the corresponding author.

The feedback form

A feedback form was developed specifically for this study that included nine questions rated on a 5 point Likert scale and four open ended questions.

Procedure

Routine individual monitoring

Teams produced consensus ratings for individuals on the GRF during regular clinical review meetings over a study period of 20 weeks. Ratings were recorded on a separate record sheet along with a note of 'any other significant events' that may have had an impact on the service user's progress. In addition, 'prediction ratings' were obtained for a sample of service users. These ratings were made at time 1 and provided 'team forecasts' of the expected level of functioning by the individual in three months' time on each domain of the GRF. The majority of individuals (n=18) were rated over a period of at least three months within the 20 week study period.

Staff feedback

At the end of the data collection period, MDT members were approached via email to complete a semi structured feedback questionnaire. The purpose of the questionnaire was fully explained, along with a statement about how the information in the questionnaire would be used and retained. Completed questionnaires were returned via email.

Data analysis

The data were collected and entered into IBM SPSS statistics 20.0 for windows (SPSS inc., 2010), which was used for all statistical and graphical analyses.

Analysis of the measurement properties of the GRF.

Donnelly et al., (2011) identified construct validity as one of the key criteria when evaluating mental health outcome measures. This was approached in two ways; first, given the differences between the three service areas, it would be expected that the tool would be able to distinguish between the functioning of the service users across the three areas. Second, the relationships between individual domains measured by the tool were investigated. Specifically, it would be anticipated that the medication and symptom management domains from the GRF would show a relationship whilst medication and substance misuse would not.

In order to assess the test-retest reliability and sensitivity to change, data from the domains obtained at different time periods were analysed. It was expected that no overall significant change would be found in the relatively short time period selected (allowing stability in measurement to be inferred) but that graphical analysis would reveal a trend towards positive change over time (allowing sensitivity to be determined).

The idiographic utility of the scale

In order to determine the interpretability and usefulness of the information generated from the GRF, a single case design was adopted (Davies & Sheldon, 2011). An individual case was selected by taking the individual with the most ratings over the longest time period (within the 20 week study) and who had a prediction rating made at time 1. The individual selected based on this criteria was typical of those within the service i.e. aged 37 with a diagnosis of schizophrenia and comorbid substance misuse and personality features. He had been in the service for 8 months at the time of the study. The data gathered were subject to visual and statistical analysis.

Usability and feasibility of the scale

Slade, Thornicroft and Glover (1999) suggest that one of the main reasons for standardised outcomes measures not being used routinely is the lack of feasibility of the tools for clinical practice. In order to establish whether the GRF would be feasible for routine use, the MDT members were asked to complete a semi structured feedback questionnaire. The results from this feedback were analysed qualitatively.

Results

Ratings

The means, standard deviation and the range of ratings used on the GRF across all of the participants (n=28) based on data from the 20 week study period are shown in Table 1. As can be seen, the mean scores are all close to '0' which is consistent with the '0' point on the GRF scale representing an expected level of functioning in service users within the service. Additionally, the full range of the GRF scale was utilised by the teams on almost all domains as indicated by the range information.

INSERT TABLE 1

ABOUT HERE

Construct validity

In order to assess the construct validity, the means and standard deviations were calculated for each of the three clinical areas. As indicated in Table 1, there is a difference in the means of the ratings provided on the GRF in the three clinical areas, with participants in the step down service generally receiving higher ratings (indicating higher functioning) across the domains and participants in the locked rehabilitation service generally receiving lowest mean ratings.

To explore whether there was a significant main effect of clinical area (step down, low secure, locked rehabilitation) on the ratings on the eight domains of the GRF, a one way between participants ANOVA was conducted with Gabriel post hoc analysis applied to determine where these differences lay. Gabriel's test was utilised because of the differences in sample sizes of the three groups. The results showed that there was a significant main effect of ratings given on the clinical area for all of the eight domains for the GRF. For medication ($F_{2,100}$ = 29.76, p<.001) all areas differed from every other, whilst significant differences were found between the locked rehabilitation ward and the step down house and between the locked rehabilitation ward and the low secure ward for symptom management ($F_{2,100}$ = 16.41, p<.001); general engagement ($F_{2,100}$ = 19.75, p<.001); relationships and support ($F_{2,100}$ = 8.47, p<.001); substance misuse ($F_{2,100}$ = 33.95, p<.001) and occupations ($F_{2,100}$ = 29.31, p<.001). Finally, there were significant differences between the step down house and the locked rehabilitation ward, and the step down house and the low secure ward for the environmental input domain ($F_{2,100}$ = 29.89, p<.001) and between step down and locked rehabilitation in relation to leave ($F_{2,100}$ = 5.86, p=.004).

To further consider the validity of the tool, two specific relationships, one were a positive relationship was expected and one where a negative relationship was expected, were examined. As anticipated, Pearson's correlation showed a significant positive relationship between the ratings on the medication domain and the symptom management domain (r= 0.60, *p* <0.01), which can be considered a large effect (Cohen, 1992). The second test revealed a significant negative relationship between the medication domain and substance misuse domain (r= -0.31, *p*<0.05); a medium negative effect (Cohen, 1992).

Stability over time and sensitivity to change

The means and standard deviations of the baseline ratings, the ratings provided at two months and the ratings at three months were calculated for 18 service users who were rated at all of these three time points. All of the ratings showed an increase in functioning over time except for substance misuse which showed a decrease in adaptive / positive functioning over time. However, repeated measures ANOVA (with adjustments where necessary due to the presence of sphericity) revealed that these changes were non-significant for all the domains.

The idiographic utility of the scale

Through the use of a single case, the ability of the GRF to record change was investigated. Person A received ratings from the low secure ward team over a period of 20 weeks. Table 2 provides the scores for Person A at each time point, the team's prediction of his likely functioning after 3 months (made at time 1) and the average low secure score for each of the domains at baseline to allow a comparison of functioning. As can be seen, in many respects, Person A's domain scores at the start are typical of those of the service users group assessed (and at or around the expected level of functioning). However, for relationships and support and for general engagement he was rated lower than his peers. Also of note is leave status which is notably higher than the group average and out of line with scores on the other domains.

INSERT TABLE 2

ABOUT HERE

The prediction rating in table 2, reveals that the team did not expect to see change at three months (T3) from the baseline score (T1) for any of the domains. However, as can be seen, four of the domains (symptom management, general engagement, relationships & support and environmental input) showed change by at least one category description over the time period. In addition, symptom management, general engagement and occupations showed (further) change between the 12 and 20 week time points (T3 to T5). Further, where change occurred it was generally gradual but could be clinically important over time (e.g. symptom management and general engagement showing change of 3 points each). The corresponding stability of domains such as medication and leave status enable inference to be drawn that these factors neither contribute to nor are affected by changes to symptom management and general engagement for this individual. In this case, his leave status was already at a very high level (+2) whilst contemporaneous notes showed engagement with psychological therapy starting immediately prior to and during the change period. This might suggest that a non-medical (in this case a psychological) intervention accounted for the symptom change. As can be seen, there are areas (namely relationships and support, and

substance misuse) in which care and treatment are still needed as the individual is below the level expected for the service (i.e. at least 0).

GRF Utility and Feasibility Feedback

Feedback from ten staff who provided completed questionnaires is summarised in Table 3. As can be seen, MDT members considered the GRF to be easy to use, with (generally) easily understood anchor descriptors. Importantly, respondents considered the GRF to be able to capture change in service users and took an appropriate amount of time to administer. However, Table 3 also suggests a mixed agreement that the frequency of reviewing service users was appropriate.   Specifically those working in the locked rehabilitation area of the service considered reviewing more frequently than monthly was not needed whereas elsewhere the 2-4 week guideline adopted was generally considered appropriate.

INSERT TABLE 3

ABOUT HERE

Feedback to the open ended questions revealed that respondents found the tool to be visually appealing and commented positively on the 'traffic light' approach to viewing the level of individual functioning.   Respondents also found the GRF to be useful in monitoring clinical progress and noted how it could be used to map onto other areas of clinical assessment i.e. the HoNOS (Wing et al., 1998) and the Mental Health Recovery Star (MHRS; MacKeith & Burns, 2008). Staff also found the process of consensus rating using the tool useful in prompting discussion. Respondents also suggested ways in which some anchor points could be revised to enhance their clarity and consistency of use.

Discussion

Measuring individual outcomes in forensic mental health services has been shown to be of great importance, however in developing new scales for outcome measurement,

investigation is needed at several steps before a measure can be adopted. This paper presents the initial stages of establishing a possible new outcome measure. This research aimed to establish whether the GRF is an effective change monitoring and outcome measure in a forensic mental health service. In relation to the research aims, the results demonstrate that the GRF shows promise, was found to be useful by the MDT and measured meaningful individual outcomes. The GRF differs from existing measures in that it is explicitly intended to promote clinical discussion, influence treatment planning and prioritisation and be a method for highlighting possible relationships between domains of functioning at the individual level.

The descriptive statistics demonstrate that the average rating given across the domains was around the central point on the GRF scale and show that the whole range of the scale was used in providing ratings across the service users. This suggests that no part of the GRF rating scale is redundant and that the complete GRF scale may be necessary to distinguish various levels of functioning within service users. This provides some initial evidence that raters' are able to distinguish between the scale points across domains offering some support for the reliability of the instrument.

As expected, the findings showed that service users in the locked rehabilitation ward generally received lower ratings across the domains of the GRF whilst those in the step down facility generally received the highest ratings. Service users who are suitable for locked rehabilitation typically have greater long term complex mental health needs than service users within a low secure service (Cope, Davernport & Maesey, 2004), but are deemed to be less risky than service users within a low secure service, as demonstrated by the greater levels of physical, relational and procedural security required by service users within a low secure mental health service. Furthermore, service users who have progressed to the low secure step down facility must have demonstrated that their mental illness symptoms and other factors related to their risk of offending have been reduced by treatment (Kennedy, 2002). Therefore these findings appear to reflect the current functioning of the service users in each clinical area, which supports the construct validity of the GRF.

The results from the analysis of the three clinical areas also demonstrated that unlike the other seven domains, service users in the step down house received the lowest ratings on the substance misuse domain (with service users within the locked rehabilitation ward receiving the highest, most positive ratings overall for this domain). This suggests that substance misuse was a greater problem for individuals in the step down house. One possible explanation for this finding is that along with moving towards recovery and discharge, service users within the step down service have more leave and greater

unaccompanied access to the community. It is possible therefore, that service users in the step down service have more opportunity and access to substances and alcohol, than those in more restricted settings.

Exploring the relationships between medication and symptom management domains revealed these to be not absolute. Whilst at the group level the correlation between these domains was positive, the single case analysis of person A indicated no apparent relationship between these two domains. Thus, although there may be an association between these domains at the group level (as would be predicted by the assumption that medication may reduce symptom experience) at the individual level a range of factors might be linked to symptom change. Examining such individual information from the domains, and their relationships, may assist formulation and future planning in relation to the maintenance of positive change.

Burgess et al., (2010) state that sensitivity to change over time is an essential criterion for an effective outcome measure. Although, it was expected that no global statistically significant changes would be identified across the domains (due to the relatively short time period of assessment), trends towards change and recovery were anticipated given the active approach to treatment provided by the service. The results supported this expectation. Such changes could be indicators of clinically meaningful change (e.g. Jacobson & Traux, 1991) as revealed across some of the domains for person A. These results suggest that the GRF can reflect clinically relevant change over time. It could be hypothesised that over a longer period of time, this trend would continue. However, this remains to be tested through the use of further cases and an extended time period.

It could be argued that the general trend towards recovery found across the domains in this study reflects a bias within the raters (i.e. they to wish to see progress). For example, Adams, Soumerai, Lomas, Ross- Degan's (1999) found that clinicians tend to overestimate performance, and suggested that social desirability bias may explain this overestimation. However, the results showing the decline in scores on the substance misuse domain over time challenges this explanation. Further, the predictions of the teams were generally for no change amongst the service user group as seen for person A. Thus recording change was against an expectation (and a public prediction) for no change to be seen. Therefore it seems reasonable to discard this explanation, at least in this study. It may be the case that the clearly defined anchor points used by the tool help mitigate such potential bias.

The single case example shows how the data can be used by clinical teams to understand possible relationships between individual domains, and to prompt consideration of possible causal relationships. Review of team discussions for Person A showed that they explained

the change in symptoms as resulting from engagement in psychological therapy whilst no change in medication had been made. In this way, recording treatment and other concurrent events alongside ratings on the domains could help teams understand what might have been the prompt for change (cf Davies, 2011).

Slade, Thornicroft and Glover (1999) suggest that one of the main reasons for standardised outcomes measures not being used routinely is the lack of feasibility of the tools. For example, research investigating the psychometric properties of the HoNOS, found that psychiatrists reported that the HoNOS was not useful when surveyed about their opinions on the HoNOS (Gilbody et al., 2002). The feedback gathered from the MDT regarding their views of the GRF, indicates that in general, it was felt that the GRF was easy to use, took an appropriate amount of time to administer and was easy to understand. There was also some agreement that the GRF is acceptable as a measure of clinical change and useful in prompting discussions around a range of areas of functioning. The majority of participants agreed that they would like to see the GRF embedded in their service. However, the feedback indicated that revising the guidance relating to the frequency of ratings might be helpful specifically, that services should determine the most appropriate frequency for rating within the guidance of 2-6 weekly. This should be based on the level of change recorded for the individual over the preceding three months and the intensity of interventions and treatment being provided. Some revisions to the anchor point descriptions for domains within the user manual were also made (clarifying terms and providing increased precision of language) based on feedback.

Limitations and Future Research Recommendations

An important limitation within this study is the short time frame (12 – 20 weeks) over which the data were collected. As an initial evaluation study, this compromise was considered acceptable to enable an assessment of the tool to be undertaken before further resources were committed. However, further research using larger samples, other settings (e.g. medium security), and over a longer time period are required. This work has begun with the continuation of the data reported here. The current approach to rating with the GRF is to use a team consensus approach. Research is required to examine the efficacy of this approach and alternatives such as single rater methods. Research is also needed to further understand the independence and relationships between the domains. Given the nature of the tool and the way it is intended for use (i.e. summary ratings based on current functioning), teams are obviously not blind to individuals and will hopefully use their summary ratings to influence the clinical approach and care accordingly (influencing

possible future ratings).   This raises some challenges for approaches to determining the reliability and validity of the tool.   For example, those making the ratings are likely to have been involved in the placement of individuals within different units thus this may effect the difference in scores by unit as seen in this study.   Further, for domains such as substance misuse, it is expected that formal mechanisms such as drug screening, would be used by the team when making their rating.   However in other domains, such 'concrete' measures may not be available and thus team discussion and observation may be the only sources of information from which to develop a rating.   Such real world challenges are important and are managed, to some extent, by the detail provided within the anchor points for ratings.   In keeping with the traditional use of goal attainment scaling, adapting the GRF to enable specific individual targets to be developed should be explored.   This could be coupled with including the individual in making the ratings or by enabling individuals to make their own ratings.   Work is underway to examine this.   Finally, the ways in which the GRF might influence and support care and decision making needs to be further understood.   Work has begun to establish a system for routinely feeding information from the ratings back to the teams to indicate changes over time.   This is coupled with work to develop a manual to detail the use and reporting of the GRF along with a video based tutorial to support this.

Conclusion

This study aimed to evaluate the GRF in terms of validity, stability, sensitivity, idiographic utility and its feasibility for use as a routine outcome measure for use in a low secure mental health setting. This study found the GRF to be a promising change monitoring and outcome measurement tool with an ability to effectively capture clinically meaningful change amongst individuals. Individual analysis showed that the GRF can enable relationships between domains to be explored and understood, as well as providing evidence of areas of progress and continued need. Feedback from those using the tool indicated that the GRF was feasible for routine use and was generally well received (subject to some minor changes). Further work to develop the materials supporting the tool and to test the tool in other settings and over a longer time period is now warranted.

References

Adams, A. S., Soumerai, S. B., Lomas, J., & Ross-Degnan, D. (1999). Evidence of self-report bias in assessing adherence to guidelines. *International Journal for Quality in Health Care*, *11*(3), 187-192. DOI: 10.1093/intqhc/11.3.187

British Psychological Society (2009) *Code of Ethics and Conduct*. Leicester: British Psychological Society

Burgess, P., Pirkis, J., Coombs, T., & Rosen, A. (2010). Review of recovery measures. *Australian Mental Health Outcomes and Classification Network*, 1-78.

Cohen, A., & Eastman, N. (2000). Needs assessment for mentally disordered offenders: measurement of 'ability to benefit' and outcome. *The British Journal of Psychiatry*, *177*(6), 493-498. DOI: 10.1192/bjp.177.6.493

Cohen, J. (1992). Statistical power analysis. *Current directions in psychological science*, *1*(3), 98-101. DOI: 10.1111/1467-8721.ep10768783

Cope, S., Davenport, S., and Measey, L. (2004). Rehabilitation and Recovery now. Council Report CR121. Royal college of psychiatrists. London. http://www.rcpsych.ac.uk/files/pdfversion/cr121.pdf . (Accessed. 13.10.14)

Cure, S. J., & Adams, C. E. (2000). Forensic trials inform the present and future. *The British Journal of Psychiatry*, *177*(2), 182-182. DOI: 10.1192/bjp.177.2.182-a

Davies, J., Howells, K., & Jones, L. (2007). Evaluating innovative treatments in forensic mental health: A role for single case methodology? *The Journal of forensic psychiatry and psychology*, *18*(3), 353-367. DOI: 10.1080/14789940701443173

Davies, J., Jones, L., & Howells, K. (2010). Evaluating individual change in forensic settings. In M. Daffern, L. F. Jones & J. Shine (Eds), *Offence Paralleling Behaviour: A case formulation approach to offender assessment and intervention* (pp. 287–302). Chichester: Wiley & Sons. DOI: 10.1002/9780470970270.ch18

Davies, J. (2011). An individual approach to assessing change. In: N. Gordon & P. Willmot (Eds), Working positively with personality disorder in secure settings: A practitioner's perspective. Chichester: Wiley & Sons

Davies, J., Maggs, R. G., & Lewis, R. (2010). The development of a UK low secure service: philosophy, training, supervision and evaluation. *International Journal of Forensic Mental Health*, *9*(4), 334-342. DOI: 10.1002/9780470973110.ch14

Davies, J. & Sheldon, K. (2011). Single case methodologies. In Kerry Sheldon, Jason Davies and Kevin Howells (Eds). *Research in Practice for Forensic Professionals*. Routledge.

Davies, J. & Maggs, R. Measuring individual change in a new low secure service: systems for individual and service evaluation. *Paper presented at International Association of Forensic Mental Health Services 9th Annual Conference,* (Edinburgh) June 2009

Davoren, M., Byrne, O., O'Connell, P., O'Neill, H., O'Reilly, K., & Kennedy, H. G. (2015a). Factors affecting length of stay in forensic hospital setting: need for therapeutic security and course of admission. *BMC psychiatry*, *15*(1), 1. DOI: 10.1186/s12888-015-0686-4

Davoren, M., Hennessy, S., Conway, C., Marrinan, S., Gill, P. & Kennedy. H.G., (2015b). Recovery and Concordance in a Secure Forensic Psychiatry Hospital – the self rated DUNDRUM-3 programme completion and DUNDRUM-4 recovery scales. BMC Psychiatry 2015, 15:61 DOI:10.1186/s12888-015-0433-x

US Department of Health and Human Services. (1999). Substance abuse and mental health services administration. *Centre for Mental Health Services, National Institutes of Health, National Institute of Mental Health*.

Donnelly, M., Scott, D., McGilloway, S., O'Neill, T., Williams, J. and Slade, M. (2011) *Patient outcomes: what are the best methods for measuring recovery from mental illness and capturing feedback from patients in order to inform service improvement?* A report commissioned by the Bamford Implementation Rapid Review Scheme. Reference: COM/4409/10; Ireland. [Online Resource] Available at: http://www.publichealth.hscni.net/sites/default/files/Patient%20Outcomes.pdf. Accessed 09/10/14

Douglas, K. S., & Kropp, R. P. (2002). A prevention based paradigm for violence risk assessment: clinical and research applications. *Criminal Justice and Behaviour,* 29: 617. DOI: 10.1177/009385402236735

Douglas, K. S., Hart, S. D., Webster, C. D., & Belfrage, H. (2013). *HCR-20v3: Assessing Risk for Violence: User Guide*. Mental Health, Law, and Policy Institute, Simon Fraser University.

Gilbody, S. M., House, A. O., & Sheldon, T. A. (2002). Psychiatrists in the UK do not use outcomes measures National survey. *The British Journal of Psychiatry*, *180*(2), 101-103. DOI: 10.1192/bjp.180.2.101

Jacobson, N. S., & Truax, P. (1991). Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *Journal of consulting and clinical psychology*, *59*(1), 12. DOI: 10.1037//0022-006x.59.1.12

Jones, J. (2007). Persons with intellectual disabilities in the criminal justice system: review of issues. *International Journal of Offender Therapy and Comparative Criminology*. 51(6), 723-733. DOI: 10.1177/0306624x07299343

Kiresuk, T. J., & Sherman, M. R. E. (1968). Goal attainment scaling: A general method for evaluating comprehensive community mental health programs. *Community Mental Health Journal*, *4*(6), 443-453. DOI: 10.1007/bf01530764

Kennedy, H. G. (2002). Therapeutic uses of security: mapping forensic mental health services by stratifying risk. *Advances in Psychiatric Treatment*, *8*(6), 433-443. DOI: 10.1192/apt.8.6.433

Kropp, P. R., & Hart, S. D. (2000). The Spousal Assault Risk Assessment (SARA) Guide: Reliability and validity in adult male offenders. *Law and Human Behaviour*, *24*, 101-118. DOI: 10.1023/a:1005430904495

MacKeith, J., Burns, S. (2008). Mental Health Recovery Star: Organisational Guide. London: Mental Health Providers Forum.

Malsch, M., and Freckelton, I. (2005) Expert bias and partisanship. *Psychological Public Policy Law 11:42*– 61. DOI: 10.1037/1076-8971.11.1.42

Morley, S., & Adams, M. (1991). Graphical analysis of single‑case time series data. *British Journal of Clinical Psychology*, *30*(2), 97-115. DOI: 10.1111/j.2044-8260.1991.tb00926.x

Nicholson, R. A. and Norwood, S. (2000). The quality of forensic psychological assessments, reports and testimony: acknowledging the gap between promise and practice. *Law and Human Behaviour 24*:9–44. DOI: 10.1023/a:1005422702678

Norusis, M. J. SPSS, Inc. (2010). *PASW statistics 18 core system user's guide. Chicago, IL: SPSS*.

Slade, M., Thornicroft, G., & Glover, G. (1999). The feasibility of routine outcome measures in mental health. *Social psychiatry and psychiatric epidemiology*, *34*(5), 243-249. DOI: 10.1007/s001270050139

Shinkfield, G., & Ogloff, J. (2014). A review and analysis of routine outcome measures for forensic mental health services. *International Journal Of Forensic Mental Health*, 13(3), 252-271. DOI:10.1080/14999013.2014.939788

Smith, A. (1994). Introduction and Overview. In Kiresuk,T., Smith, A., and Cardillo, J. E. (Eds.) *Goal attainment scaling: Applications, theory and measurement.* Lawrence Erlbaum Associates Inc. Hillside, New Jersey

Stein, G. S. (1999). Usefulness of the Health of the Nation Outcome Scales. *British Journal of Psychiatry, 174*, 375 –377 DOI: 10.1192/bjp.174.5.375

Tashakkori, A., & Teddlie, C. (1998). *Mixed methodology: Combining qualitative and quantitative approaches* (Vol. 46). SAGE Publications, Incorporated.

Thornley, B., & Adams, C. (1998). Content and quality of 2000 controlled trials in schizophrenia over 50 years. *British Medical Journal, 317* (7167), 1181-1184. DOI: 10.1136/bmj.317.7167.1181

Webster, C. D., Douglas, K. S., Eaves, D., & Hart, S. D. (1997). HCR–20: Assessing risk for violence (Version 2). Vancouver, BC, Canada: *Mental Health, Law, and Policy Institute*, Simon Fraser University

Wettstein, R. M. (2005). Quality and Quality Improvement in Forensic Mental Health Evaluations. *Journal of the American Academy of Psychiatry and the Law. 33:*158–75.

Wing, J.K., Beevor, A.S., Park, R.H., Haddon, S.B.J., Burns, S. & Burns, A. (1998) Health of the nation outcome scales (HoNOS) research and development. *British Journal of Psychiatry*; 172, pp11-18. DOI: 10.1192/bjp.172.1.11

Yiend, J., Chambers, J. C., Burns, T., Doll, H., Fazel, S., Kaur A., Sutton, L., and Fitzpatrick, R. (2011). Outcome measurement in forensic mental health research: an evaluation. *Psychology, Crime and Law, 17:3,* 277-29. DOI: 10.1080/10683160903203938

	Ratings from all participants across all time points				Ratings by clinical			
					Step down house (n= 26)		Low secure ward (n= 34)	
GRF domain	M	SD	MIN	MAX	M	SD	M	SD
Medication*	-0.10	0.92	-3	2	0.73	0.87	0.00	0.00
Symptom management*	0.34	1.52	-3	3	1.15	1.62	0.82	1.29
General engagement*	-0.50	1.57	-3	3	-0.19	1.70	0.44	1.40
Leave^	0.81	1.38	-3	3	1.46	1.80	0.88	1.45
Relationships & support*	-0.85	0.96	-3	2	-0.27	1.25	-0.76	0.70
Substance Misuse*	-0.26	2.06	-3	3	-1.42	0.95	-1.32	0.73
Occupations*	-0.25	1.35	-3	2	0.58	1.33	0.32	1.04
Environmental input*	0.17	1.34	-3	3	1.58	1.47	-0.32	0.81

Table 1: The means and standard deviations and range for the total ratings on each domain and rati
for each of the domains of the GRF

Differences between areas examined using one way ANOVA; * = significant at $p<0.01$; ^ = significant at $p<0$

	Low secure baseline rating		Person A			
GRF domain	M	SD	Time 1	Time 2	Time 3	Time 4
Medication	0.00	0.00	0	0	0	0
Symptom management	0.62	1.30	-1	1	1	2
General engagement	0.00	1.70	-2	-1	0	1
Leave	0.00	1.70	2	2	2	2
Relationships &support	-0.75	1.03	-2	-2	-1	-1
Substance Misuse	-1.12	0.35	-1	-1	-1	-1
Occupations	0.12	1.25	0	-1	0	1
Environmental input	-0.62	0.74	-1	0	0	0

Table 2: Ratings for Person A over a 20 week period with prediction rating and low secure baseline comparison.

	Strongly disagree	Disagree	Neither disagree or agree	Agree	Strongly agree
The Global Review Form is easy to use	0	0	0	8	2
The Global Review Form takes an appropriate amount of time to administer	0	0	0	8	2
The Global Review Form captures change in presentation of service users within your service	0	0	0	8	2
The time period between using the Global Review Form to assess the service users progress has been appropriate	0	2	2	4	2
The anchor point descriptions of the Global Review Form are easy to understand	0	1	0	9	0
The Global Review Form is useful for routine monitoring of a service user's progress in your setting	0	0	1	8	1
The Global Review Form is a practical way of objectively monitoring and reviewing service users within your setting	0	0	1	9	0
The Global review form is useful in assisting with clinical decision making	0	1	3	6	0
The Global Review Form adequately captures the overall presentation of the service users you work with	0	0	4	6	0

Table 3: Number of MDT members endorsing each feedback statement.