Psychiatry Research 243 (2016) 312–317

Contents lists available at ScienceDirect

Psychiatry Research

journal homepage: www.elsevier.com/locate/psychres

The reliability of the Personal and Social Performance scale – informing its training and use

Sarah White [a,*], Christianne Dominise [a], Dhruv Naik [a], Helen Killaspy [b]

[a] Population Health Research Institute, St George's, University of London, Cranmer Terrace, London SW17 0RE, UK
[b] Division of Psychiatry, University College London, Maple House, 149 Tottenham Court Road, London W1T 7NF, UK

ARTICLE INFO

Article history:
Received 15 January 2016
Received in revised form
21 June 2016
Accepted 26 June 2016
Available online 29 June 2016

Keywords:
Inter-rater reliability
Social functioning
Measurement scale

ABSTRACT

Social functioning is as an important outcome in studies of people with schizophrenia. Most measures of social function include a person's ability to manage everyday activities as well as their abilities to engage in leisure and occupational activities. The Personal Social Performance (PSP) scale assesses functioning across four dimensions (socially useful activities, personal and social relationships, self-care, disturbing and aggressive behaviours) rather than one global score and thus has been reported to be easier to use. In a pan-European study of people with severe mental illness a team of 26 researchers received training in rating the scale, after which the inter-rater reliability (IRR) was assessed and found to be not sufficiently high. A brief survey of the researchers elicited information with which to explore the low IRR and their experience of using the PSP. Clinicians were found to have higher IRR, in particular, psychologists. Patients' employment status was found to be the most important predictor of PSP. Researchers used multiple sources of information when rating the scale. Sufficient training is required to ensure IRR, particularly for non-clinical researchers, if the PSP is to be established as a reliable research tool.

1. Introduction

Many studies have reported deficit in social functioning as a core feature of people suffering from schizophrenia (Bellack et al., 2007; Dickerson et al., 1999) and social functioning is therefore recognised as an important outcome in studies of this group, (Burns and Patrick, 2007). The concept of social functioning usually includes the ability of a person to function in different personal and societal roles and their satisfaction with their ability to meet these roles. Most measures of social function include a person's ability to manage everyday activities (such as self-care, shopping, cooking, cleaning and budgeting) as well as their abilities to engage in leisure and occupational activities (Mueser and Tarrier, 1998). A limitation of social functioning rating scales is the lack of consistency in the inclusion of objective indicators (e.g. employment, having a partner, living independently) and subjective indicators (e.g. self-rated wellbeing and views on their social situation) (Apiquian et al., 2009).

The most widely used scale of social functioning in people with severe mental illness is the Global Assessment of Functioning (GAF) (American Psychiatric Association, 1987), a revised version of the Global Assessment Scale (Endicott et al., 1976). It has been used as a clinical assessment tool as well as an outcome measure in research, with data being aggregated at the individual or group/sample level. The GAF includes assessment of three dimensions of functioning; social, occupational, and psychological symptoms, but the rater makes an overall single rating between 0 and 100, where 100 is the highest level of social function.

The Personal and Social Performance scale (PSP) (Morosini et al., 2000) is a revision of the Social and Occupational Functioning Assessment Scale (SOFAS) (Nietzel and Wakefield, 1996). The SOFAS was included in DSM-IV and is similar to GAF but only rates social and occupational functioning rather than symptoms. The main advantage of the PSP over GAF and SOFAS is that it assesses functioning in four dimensions (socially useful activities, personal and social relationships, self-care, disturbing and aggressive behaviours) rather than one global score (Juckel et al., 2008) and thus has been reported to be easier to use (Burns and Patrick, 2007). Clinicians with any level of experience and from different professional backgrounds can easily be trained to use the PSP (Morosini et al., 2000). However, like the GAF, the PSP's main limitation is that it is rated on the basis of clinical information about the person, obtained from the person themselves, clinical staff and case notes, rather than through a structured interview. Obtaining access to relevant information can therefore pose a difficulty in its use (Nasrallah et al., 2008).

A pan-European study of people with severe mental illness living in longer term rehabilitative settings (the "DEMoBinc study")

* Corresponding author.
E-mail address: swhite@sgul.ac.uk (S. White).

http://dx.doi.org/10.1016/j.psychres.2016.06.047
0165-1781/

required a measure of social functioning in order to assess the range of functioning across the large, possibly heterogeneous sample (Killaspy et al., 2009). The GAF was chosen as a commonly used, relevant measure. The PSP was also proposed as a newer measure which may provide a more rounded assessment of social functioning. A team of 26 researchers from the 10 countries participating in the study received training in rating both the GAF and PSP, after which the inter-rater reliability (IRR) of both measures were assessed based on the ratings of 10 clinical vignettes. The IRR of GAF was high (intra-class cluster coefficient (ICC)=0.88, 95% confidence interval (CI): 0.76, 0.96) but considerably lower for the PSP (ICC=0.64, 95% CI: 0.44, 0.86). Both measures were subsequently used in a cross sectional study of 1750 patients, GAF being reported in the primary analysis (Killaspy et al., 2012) because of its more acceptable IRR.

This paper reports two related post-hoc analyses which were conducted to provide possible explanations for the poor IRR of the PSP and to answer the following research questions:

1) What rater characteristics are associated with varying inter-rater reliability of the PSP?

2) What patient characteristics are taken into account when rating the PSP?

2. Methods

2.1. Study design

The DEMoBInc researchers were contacted and asked to complete a brief survey about their professional background, length of experience working in mental health services and their experience of rating the PSP and GAF during the DEMoBinc research interviews. These data were used to perform two analyses. The first investigated the characteristics of the researchers and PSP vignette ratings to establish whether rater characteristics could explain variability in IRR. The second investigated which patient variables (assessed in the DEMoBinc research interview) were considered by the researchers to be most useful in informing their rating of the PSP and GAF, particularly exploring whether different information was used to complete the two scales.

2.2. Procedures

As part of the DEMoBInc study, 1750 service users of 213 longer term mental health rehabilitation units across ten European countries were interviewed. For details of selection and recruitment see Killaspy et al. (2012). Characteristics of the service users have also been previously published (Killaspy et al., 2012) but in summary the 1750 service users were recruited from 2495 approached (70% response rate). The mean age was 46 years (range to 18–87 years) with 62% male.

The interview comprised assessments of the service user's i) experience of care (Your Treatment and Care (Webb et al., 2000), ii) autonomy (Resident Choice Scale (Hatton et al., 2004), iii) quality of life (Manchester Short Assessment of Quality of Life (Priebe et al., 1999), iv) rating of the service's therapeutic environment (Good Milieu Index (Røssberg and Friis, 2003), v) use of services over the previous six months (Client Services Receipt Inventory (Beecham and Knapp, 2001)) and sociodemographic characteristics. Data were also collected on markers of recovery, such as participation in voting in the last election, having a bank account, being in charge of their own finances, and negative experiences within the unit in the last year (e.g. being shouted at, frightened or threatened, and/or being physically or sexually abused). At the end of the interview researchers made ratings of the service user's social functioning using GAF and PSP.

All 26 DEMoBinc researchers were trained in the use of the service user interview materials (including GAF and PSP) by senior research team members at an extended research team meeting in February 2009. The GAF and PSP training workshop consisted of trainers introducing and explaining the two measures to the researchers and demonstrating their use. The researchers were then asked to complete GAF and PSP ratings of a series of training vignettes. The ratings were compared and discussed, exploring and resolving discrepancies to achieve agreement. At the end of the training session, the researchers were asked to provide GAF and PSP ratings of ten further clinical vignettes. All ratings were collated and entered into an Excel spreadsheet for future analysis. When researchers rated the PSP they scored each of the four domains of the PSP using the six level categorical responses available. The four domains are A) socially useful activities, including work and study; B) personal and social relationships; C) self-care; and D) disturbing and aggressive behaviours. These ratings were subsequently converted into 10 point band scores for each vignette using the published guidance (Morosini et al., 2000) by SW. This meant that no overall PSP rating between 1 and 100 was made. The 10 point band ratings are analysed further in this study.

In February 2012 a questionnaire was sent to all the DEMoBinc researchers by email, along with a copy of the original research interview schedule used in the DEMoBinc study. Researchers were asked for the following information: their age; gender; professional training (categorised as medicine, psychology, other science, non-science); current profession (whether they considered themselves to be mainly a researcher or mainly a clinician); current occupation (categorised as psychiatrist, psychologist or other); and years working in mental health (categorised as 0–5 years, 6–10 years, more than 10 years); which components of the PSP they had found most difficult to rate; the features of the PSP they felt were most likely to lead to inconsistency in ratings; which questions within the DEMoBinc research interview had provided the most useful information for rating the GAF and the PSP; whether any of their own observations of the service users (e.g. appearance, communication skills) had influenced their rating of the GAF and PSP; whether they had sought additional information from other sources (medical records, clinical staff) to inform their rating of the GAF and PSP; and if there was other information they would have liked but were unable to access to inform their ratings. Finally they were asked if they had used the GAF or PSP prior to the DEMoBInc study. Weekly email reminders were sent to the researchers over a period of one month to maximise response.

2.3. Data analysis

2.3.1. Rater characteristics and inter-rater reliability of the PSP

Inter rater reliability (the level of agreement between raters) was calculated using intra class coefficients (ICC). The specific type of ICCs calculated here resulted from a two-way mixed analysis of variance where absolute agreement between raters is integral and needs to be generalised to the case of a single measure (McGraw and Wong, 1996). Cicchetti (1994) presents cut-offs to be applied to ICCs in order to give qualitative descriptions of the degree of agreement; ICC values are deemed ‘excellent’ if greater than or equal to 0.75, ‘good’ if between 0.6 and 0.74, ‘fair’ if between 0.4 and 0.59, ‘poor’ if below 0.4.

2.3.2. Patient characteristics and rating the PSP

The DEMoBinc research interview questions that were identified by at least 50% of the researchers as being useful in their ratings of either GAF or PSP were summarised using frequencies and percentages for categorical variables and mean, standard deviation, minimum and maximum values for interval variables. This

set of variables was entered into a multiple regression model as independent variables, with GAF as the dependent variable, then PSP. The regression models were then refitted omitting all variables which were not statistically significant at the 5% level. This produced two distinct models indicating which of the independent variables were associated with GAF and PSP. The unstandardised regression coefficients (*B*) and 95% CIs are presented as well as the standardised coefficients (*Beta*) to aid the assessment of the relative importance of the independent variables in the two models. The assumptions made in these regression models were tested by examination of collinearity statistics and analysis of residuals.

All statistical analyses were conducted using IBM SPSS Statistics v22 (IBMCorp, 2013).

3. Results

3.1. Rater characteristics and inter-rater reliability of the PSP

Of the 26 individuals contacted 58% ($n=15$) replied with a completed questionnaire. The mean age of respondents was 35.1 years (range 27–63) and most were female (11, 73%). Six (40%) of the respondents' first degree was in the field of psychology and four (27%) had studied medicine. Ten (60%) responders described themselves as mainly researchers and five (40%) as mainly clinicians. Seven (46%) had been working in the mental health field for five or less years and four (27%) for more than 10 years. Six of the 14 (43%) respondents had used GAF before and only one had used the PSP before the DEMoBinc study.

The IRR was recalculated using the vignette ratings of the survey respondents only. The ICC for the respondents was 0.78 (95% CI 0.61, 0.92). In examining the vignette ratings, the 41–50 and 51–60 ranges were most commonly used and the 1–10 and 11–20 ranges were not used at all. All but one vignette received ratings in three adjacent range bands. Vignette 5 had the greatest spread in ratings, across the four 10 point bands between 41–50 and 71–80.

All ICCs presented in Table 1 were above 0.75 apart from one, indicating excellent reliability. There was a higher level of agreement of ratings amongst those who described themselves as mainly clinicians compared to those who described themselves as mainly researchers. Inter-rater reliability was highest amongst psychologists. The ICCs appeared to vary depending on the length of time raters had spent working in mental health services; those with the least experience in the field had greater reliability than those with 6–10 years of experience, but the group with more than 10 years of experience had the highest ICC (0.9).

One respondent felt that raters with clinical training would be more consistent. Suggestions given on how to improve the training they had received on rating the PSP were; more examples to make the distinction between adjacent severity ratings clearer, using video-taped interviews for practice ratings rather than vignettes and, in line with this, emphasising the need to take into account visual cues during an interview. This latter point was felt to be particularly important for non-clinicians.

Table 1
Inter-rater reliability of the PSP by rater characteristics.

Variable	Label	*n* (%)	ICC	95% CI
Current occupation	**Researcher**	9 (60%)	0.77	0.58, 0.92
	Clinician	6 (40%)	0.82	0.64, 0.94
Professional background	**Psychiatrist**	4 (27%)	0.79	0.56, 0.93
	Psychologist	5 (33%)	0.82	0.64, 0.94
	Other	6 (40%)	0.78	0.55, 0.93
Years working in mental health	**0–5**	7 (46%)	0.79	0.60, 0.93
	6–10	4 (27%)	0.68	0.38, 0.89
	More than 10	4 (27%)	0.90	0.78, 0.97

Table 2
Items of the DEMoBInC study research interview identified by > 50% of the sample as informing their rating of the GAF and PSP.

Questions	GAF $n=12$	PSP $n=11$
MANSA		
Employment status	11 (92%)	11 (100%)
What is your occupation (if employed)?	6 (50%)	9 (82%)
How many hours a week do you work (if employed)?	10 (83%)	10 (91%)
How many children do you have?	1 (8%)	7 (64%)
Do you have anyone who you would call a close friend (includes family if subject prefers but not professionals)?	8 (67%)	10 (91%)
Have you seen a friend in the last week?	6 (50%)	7 (64%)
How satisfied are you with the number and quality of your friendships?	7 (58%)	10 (91%)
How satisfied are you with your leisure activities?	7 (58%)	8 (73%)
In the past 2 weeks, have you been out to play or watch a sport?	7 (58%)	5 (45%)
In the past 2 weeks, have you been out shopping?	7 (58%)	6 (55%)
In the past 2 weeks, have you been for a ride in a bus, car or train other than for transport to and from work?	7 (58%)	5 (45%)
In the past year, have there been times when you would have liked to have had more leisure activity but were unable?	5 (42%)	7 (64%)
In the past year have you been accused of a crime?	2 (17%)	8 (73%)
How satisfied are you with the people that you live with?	1 (8%)	8 (73%)
How satisfied are you with your sex life?	4 (33%)	6 (55%)
How satisfied are you with your relationship with your family?	6 (50%)	8 (73%)
Markers of Recovery		
Do you have a bank (or post office) account?	6 (50%)	4 (36%)
Do you have charge of all your finances?	8 (67%)	6 (55%)
Resident Choice Scale		
How much choice do you have in this area of your life, having a partner?	4 (33%)	6 (55%)

3.2. Patient characteristics and rating the PSP

One of the respondents who had completed the PSP training had not gone on to do the research interviews so this second analysis is based on 14 of the 26 (54%) DEMoBinc researchers. Twelve respondents identified a total of 98 items from the DEMoBinc research interview that they found useful in rating the GAF and PSP, of which 19 were identified by over 50% of respondents (Table 2). The questions most commonly identified related to employment, leisure activities and personal relationships. Some variables were not considered useful in rating both GAF and PSP; the 'number of children', 'being accused of a crime' and 'satisfaction with the people they live with' were noted as useful for PSP rather than GAF ratings.

Two of the 19 variables identified by respondents were not included in the multiple regression analyses; the occupation variable provided only qualitative data and the hours of work variable only applied to those that had a job and thus was highly skewed (only 15% non-zero), thus contributing little more information than the categorical employment status variable.

Of the 17 variables entered into the preliminary regression model, eight were found to be significant predictors of GAF and PSP and the final models with these eight variables are shown in Table 3. Seven variables were found to be predictive of both GAF and PSP: employment status; having a close friend; going out shopping; use of transport; a desire to do more leisure activities; being satisfied with the people they live with; having charge of

Table 3
Multiple regression analysis showing relationship of identified variables with GAF and PSP.

Variables ($n=1750$)	**Summary statistics** *n* (%)	**GAF** *B* (95% *CI*) *Beta* (*p-value*)	**PSP** *B* (95% *CI*) *Beta* (*p*-value)
Socio-demographic variables			
In paid/sheltered employment or training/education *Yes*	261 (15%)	9.8 (8.0, 11.7) 0.25 (< 0.001)	1.1 (0.9, 1.3) 0.26 (< 0.001)
MANSA items			
Do you have anyone you would call a close friend, can be family but not professionals? *Yes*	1147 (66%)	2.8 (1.4, 4.2) 0.09 (< 0.001)	0.4 (0.2, 0.5) 0.12 (< 0.001)
In the past 2 weeks, have you been out to play or watch a sport? *Yes* ($n=1742$)	502 (29%)	1.9 (0.4, 3.4) 0.06 (0,013)	
In the past 2 weeks, have you been out shopping? *Yes* ($n=1741$)	1216 (70%)	−4.7 (−6.2, −3.2) −0.15 (< 0.001)	−0.3 (−0.5, −0.2) −0.1 (< 0.001)
In the past 2 weeks, have you ridden in a bus, car or train other than for work? *Yes* ($n=1742$)	1008 (58%)	3.4 (2.0, 4.8) 0.11 (< 0.001)	0.3 (0.2, 0.5) 0.11 (< 0.001)
In the past year, would you have liked more leisure activity but were unable? *Yes* ($n=1599$)	840 (48%)	−1.9 (−3.2, −0.5) −0.07 (0.006)	−0.3 (−0.4, −0.1) −0.09 (0.003)
How satisfied are you with the people that you live with? *Mean* (*SD*) *Min-Max*	4.97 (1.382) 1–7	0.5 (0.0, 1.0) 0.05 (0.035)	0.05 (0.0, 1.0) 0.05 (0.032)
Markers of recovery			
Do you have charge of all your finances? *Yes* ($n=1749$)	693 (40%)	4.8 (3.4, 6.2) 0.16 (< 0.001)	0.5 (0.3, 0.6) 0.15 (< 0.001)
Resident Choice Scale			
How much choice do you have in this area of your life, having a partner? *Mean* (*SD*) *Min-Max*	2.9 (1.3) 1–4		0.05 (−0.01, 0.1) 0.04 (0.074)

their finances. The standardised regression coefficients were very similar, the greatest difference between them just 0.05 for going out shopping. With regard to the standardised regression coefficients, employment status was the variable most associated with GAF and PSP rating, with effect sizes of 0.25 and 0.26 respectively. The unstandardized regression coefficients indicated that people who were in some form of paid employment were rated approximately 10 points higher on GAF and nearly one 10 point band higher on PSP than those who were unemployed. Curiously, going out shopping was negatively associated with both GAF and PSP with small but significant effect sizes, −0.15 and −0.1 respectively. People who would have liked to do more leisure activities but were unable were also rated lower, with negative effect sizes of −0.07 and −0.09 for GAF and PSP respectively.

With respect to the variables that were not associated with both GAF and PSP, going out to play or watch sport was significantly ($p=0.013$) and positively associated with GAF (Beta=0.06). The 'having a partner' item of the Resident Choice Scale was significantly associated with PSP in the preliminary regression model but was no longer significant at the 5% level in the final model ($p=0.074$). Collinearity statistics and analysis of residuals indicated that the assumptions made in regression modelling were not violated. The level of variability of the two measures was also examined using the coefficient of variation (CV) statistic. The CVs for GAF and PSP were low and very similar, at 31% and 27% respectively, indicating similar variation in the two measures in this sample.

When asked about any other factors they took into account when rating the two measures, nine out of 14 (64%) of the researchers reported that the patient's appearance (such as their clothing, physical appearance and hygiene) influenced their ratings of GAF and PSP, respectively. The person's ability to communicate was also considered an important factor in rating GAF by 10 out of 14 (71%) respondents, but only 5 (36%) considered it important for rating PSP.

Most of the researchers (71%) spoke with clinical staff to gain additional information to aid with their GAF and PSP scoring. Only a few suggested other sources they thought would be useful, including having more time to talk to staff, having access to patients' case notes and speaking to relatives.

Survey respondents reported various difficulties in rating the PSP. These included the wide range of activities included in the 'socially useful activities' domain (domain A). Confusion was expressed as to how to rate a patient who wasn't able to undertake any work, paid or voluntary, training or education, but was able to manage their own housework. Another example was given describing two patients who were both able to conduct the same activity and were therefore rated the same, yet one patient was an inpatient, the other lived independently. The inference was that a person able to live independently would be likely to have a higher level of social function than an inpatient. The domain 'personal and social relationships' (Domain B), was also considered difficult to rate as researchers found it hard to observe these behaviours and sensitive to discuss. There was also confusion about how to rate relationships with people the patient rarely had contact with but where the relationship was deemed to be important by the patient.

4. Discussion

4.1. Discussion of findings

This study was prompted by our finding that the IRR of the PSP was lower than the GAF when rated by a large group of trained researchers from different countries. When exploring the variation in ICCs between different groups of researchers some interesting results came to light. Whilst the majority of ICCs were within the 'excellent' category those respondents considering themselves primarily clinicians had greater IRR than those who considered themselves solely researchers and psychologists had the highest agreement compared to other professional disciplines. In a Mexican study where all raters were clinicians, that validated the scale for use with adolescents (Ulloa et al., 2015), very high ICCs were found for all domains (> 0.82) and for the total score (0.85). In the original validation of the scale (Morosini et al., 2000) an overall ICC of 0.98 was found but non-qualified staff (nursing aides) were less consistent in their ratings, both amongst themselves and compared with other professional groups. Patrick and colleagues (2009) tested the IRR of the PSP on a group of clinically trained raters using a similar set of vignettes as used in our study. An ICC of 0.9 was found when calculating the statistic based on the 10 point band PSP ratings (Patrick et al., 2009). It therefore appears that having a clinical training is associated with better IRR and this

may explain some of the variance in our findings.

We also found variation in the IRR of the PSP with respect to raters' length of experience in the mental health field. One possible explanation is that those with little experience paid closer attention to the training than those with moderate experience and thus had greater IRR. However, those with more extensive experience may have had a much deeper understanding of social functioning and therefore had excellent IRR.

The manner in which the PSP is scored means that a one point difference in how a rater rates just one of the four dimensions can take the overall rating into an adjacent 10 point band. This is particularly the case for Domain D (disturbing and aggressive behaviours). Therefore a small error rating this dimension in particular impacts significantly on the overall PSP score. This could explain the relatively wide range of scores we found for many of the vignettes.

When asked about their experience of using the GAF and PSP during the course of the DEMoBinc study, researchers reported more use of observations of the patient's presentation and communication skills when rating GAF than PSP. This may be due to the different structures of the scales, where the GAF score is assigned as an overall rating of functioning whereas the PSP requires ratings of the four dimensions which then dictate an overall rating within a ten point band. In other words, the PSP has a more structured and more restrictive means of rating. Broadly, researchers reported using similar elements of the research interview to inform their ratings. The following variables were reported by researchers and found to be statistically associated with ratings of both PSP and GAF: employment status; having a close friend; going out shopping; use of transport; a desire to do more leisure activities; being satisfied with the people they live with; having charge of their finances. Going out shopping in the last two weeks was negatively associated with both GAF and PSP which appears a counterintuitive finding. A possible explanation of this is that researchers did not see 'going out shopping' as a socially useful activity and rated patients lower as a consequence. This is potentially an issue which could be explored during training.

Researchers also asked staff and consulted patient case notes for additional information where necessary. One small study investigated the IRR of the GAF amongst three clinically trained raters when completed using only psychiatric case records (Mirandola et al., 2000); very high levels of IRR were found.

The set of vignettes generally used to test the IRR of a team of researchers after PSP training are those developed for the purpose by the scale authors (Morosini et al., 2000). In this study these were used as were a further six written by experienced academic clinicians. The content of vignettes used for training and checking IRR need to be considered and additional modes of training may be necessary. The ideal scenario in which to test IRR would be by using actual subjects but in the case of this large, multi-country, team of researchers this was not feasible.

4.2. Limitations

The ICC for the survey respondents was 10% higher than for all researchers who were trained. This suggests that the respondents were more homogeneous in their characteristics than the whole sample of researchers who underwent the training. The IRR of the GAF and PSP was formally tested directly after the research team had been trained so should have been at its maximum. The IRR was not tested again over the course of data collection. However the ten sites interviewed over 150 service users each over the following eight months, a very intense recruitment period, so the skills gained from the training would have been utilised and built upon over that recruitment period.

The response rate for the follow-up survey was moderate as some researchers had completed their contracted posts and were no longer contactable. Although the research team was relatively large the moderate response rate means that the inter-rater reliability analyses is based on a small sample. Time had obviously passed between the original training, carrying out the DEMoBinc study research interviews and completing our surveys for this study. While this would not have affected the accuracy of the data on the characteristics of the researchers who responded, their recollection of their experiences of the training and of undertaking the research interviews may have been subject to recall bias.

4.3. Conclusions

This study explored the possible reasons for variability in the reliability of ratings of the PSP amongst a large team of researchers from different countries with varying clinical and research experience. Our findings confirm those of other studies, in that clinical experience appears to be associated with greater IRR. We also identified eight specific items from other standardised measures that are useful in informing the ratings of social function using PSP and GAF. Whilst the scale may provide a more comprehensive measure of social functioning than other commonly used measures, researchers must be appropriately trained in its use and encouraged to corroborate their clinical observations with staff reports and information from case records.

Acknowledgements

This paper uses data collected as part of the DEMoBinc study which was funded by a three-year grant from the European Commission's 6th Framework (SP5A–CT–2007-044088). The DeMoBinc group are Professor Helen Killaspy, Dr Christine Wright, Professor Dr. Thomas Kallert, Professor Jorge Cervilla, Professor Jiri Raboch, Dr Georgi Onchev, Ass. Professor Giuseppe Dell'Acqua, Professor Durk Wiersma, Professor Andrzej Kiejna, Ass. Professor Dimitris Ploumpidis, Professor Jose Caldas de Almeida, Professor Michael King, Dr Sarah White, Professor Paul McCrone. We would like to thank members of the research team who responded to the survey for their thoughtful responses.

References

American Psychiatric Association, 1987. Diagnostic and Statistical Manual of Mental Disorders, 3rd Ed., revised, Author, Washington, DC.

Apiquian, R., Ulloa, R.E., Herrera-Estrella, M., Moreno-Gómez, A., Erosa, S., Contreras, V., Nicolini, H., 2009. Validity of the Spanish version of the Personal and Social Performance Scale in schizophrenia. Schizophr. Res. 112 (1), 181–186.

Beecham, J., Knapp, M., 2001. Costing psychiatric interventions. In: Thornicroft, G. (Ed.), Measuring Mental Health Needs. Gaskell, London, pp. 200–224.

Bellack, A.S., Green, M.F., Cook, J.A., Fenton, W., Harvey, P.D., Heaton, R.K., Patterson, T.L., et al., 2007. Assessment of community functioning in people with schizophrenia and other severe mental illnesses: a white paper based on an NIMH-sponsored workshop. Schizophrenia bulletin 33 (3), 805–822.

Burns, T., Patrick, D., 2007. Social functioning as an outcome measure in schizophrenia studies. Acta Psychiatr. Scand. 116 (6), 403–418.

Cicchetti, D.V., 1994. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol. Assess. 6 (4), 284–290.

Dickerson, F., Boronow, J.J., Ringel, N., Parente, F., et al., 1999. Social functioning and neurocognitive deficits in outpatients with schizophrenia: a 2-year follow-up. Schizophrenia Res 37 (1), 13–20.

Endicott, J., Spitzer, R.L., Fleiss, J.L., Cohen, J., 1976. The Global assessment scale: a procedure for measuring overall severity of psychiatric disturbance. Arch. Gen. Psychatry 33 (6), 766–771.

Hatton, C., Emerson, E., Robertson, J., Gregory, N., Kessissoglou, S., Walsh, P.N., 2004. The Resident choice scale: a measure to assess opportunities for self-determination in residential settings. J. Intellect. Disabil. Res. 48, 103–113.

IBM Corp. Released, 2013. IBM SPSS Statistics for Windows, Version 22.0., IBM Corp, Armonk, NY.

Juckel, G., Schaub, D., Fuchs, N., Naumann, U., Uhl, I., Witthaus, H., Hargarter, L.,

Bierhoff, M., Brüne, M., 2008. Validation of the Personal and Social Performance (PSP) scale in a German sample of acutely ill patients with schizophrenia. Schizophr. Res. 104 (1), 287–293.
Killaspy, H., King, M., Wright, C., White, S., McCrone, P., Kallert, T., Cervilla, J., Raboch, J., Onchev, G., Mezzina, R., Kiejna, A., Ploumpidis, D., Caldas de Almeida, J., 2009. Study protocol for the development of a European measure of best practice for people with long term mental illness in institutional care (DEMo-Binc). BMC Psychiatry 9, 36.
Killaspy, H., White, S., Wright, C., Taylor, T.L., Turton, P., Kallert, T., Schuster, M., Cervilla, J.A., Brangier, P., Raboch, J., Kalisova, L., 2012. Quality of longer term mental health facilities in Europe: validation of the quality indicator for rehabilitative care against service users' views. PLoS ONE 7 (6), e38070, e38070.
McGraw, K.O., Wong, S.P., 1996. Forming inferences about some intraclass correlation coefficients. Psychol. Methods 1 (1), 30–46.
Mirandola, M., Baldassari, E., Beneduce, R., Italo, A., Segala, M., Tansella, M., 2000. A standardized and reliable method to apply the Global Assessment of Functioning (GAF) scale to psychiatric case records. Int. J. Methods Psychiatr. Res. 9 (2), 79–86.
Morosini, P.L., Magliano, L., Brambilla, L., Ugolini, S., Pioli, R., 2000. Development, reliability and acceptability of a new version of the DSM-IV Social and Occupational Functioning Assessment Scale (SOFAS) to assess routine social functioning. Acta Psychiatr. Scand. 101 (4), 323–329.
Mueser, K.T.E., Tarrier, N.E., 1998. Handbook of Social Functioning in Schizophrenia. Allyn & Bacon, Needham Heights, Massachusetts.
Nasrallah, H., Morosini, P., Gagnon, D.D., 2008. Reliability, validity and ability to detect change of the Personal and Social Performance scale in patients with stable schizophrenia. Psychiatr. Res. 161 (2), 213–224.
Nietzel, M.T., Wakefield, J.C., 1996. American psychiatric association diagnostic and statistical manual of mental disorders. Contemp. Psychol. 41, 642–651.
Patrick, D.L., Burns, T., Morosini, P., Rothman, M., Gagnon, D.D., Wild, D., Adriaenssen, I., 2009. Reliability, validity and ability to detect change of the clinician-rated Personal and Social Performance scale in patients with acute symptoms of schizophrenia. Curr. Med. Res. Opin. 25 (2), 325–338.
Priebe, S., Huxley, P., Knight, S., Evans, S., 1999. Application and results of the manchester short assessment of quality of life (MANSA). Int. J. Soc. Psychiatr. 45, 7–12.
Røssberg, J.I., Friis, S., 2003. A suggested revision of the ward atmosphere scale. Acta Psychiatr. Scand. 108, 374–380.
Ulloa, R.E., Apiquian, R., Victoria, G., Arce, S., González, N., Palacios, L., 2015. Validity and reliability of the Spanish version of the Personal and Social Performance scale in adolescents with schizophrenia. Schizophr. Res. 164 (1), 176–180.
Webb, Y., Clifford, P., Fowler, V., Morgan, C., Hanson, M., 2000. Comparing patients' experience of mental health services in England: a five-trust survey. Int. J. Health Care Qual. Assur. 13 (6), 273–281.