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