Psychiatry Research 188 (2011) 446–452

Contents lists available at ScienceDirect

Psychiatry Research

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

Recommendations to improve the Positive and Negative Syndrome Scale (PANSS) based on item response theory

Stephen Z. Levine [a,*], Jonathan Rabinowitz [a], Dimitris Rizopoulos [b]

[a] Bar Ilan University, Ramat Gan 52900, Israel
[b] Department of Biostatistics, Erasmus Medical Center, Rotterdam, The Netherlands

ARTICLE INFO

Article history:
Received 27 June 2010
Received in revised form 15 November 2010
Accepted 11 March 2011

Keywords:
Schizophrenia
Severity of Illness Index
Psychometrics
PANSS
Item Response Theory

ABSTRACT

The adequacy of the Positive and Negative Syndrome Scale (PANSS) items in measuring symptom severity in schizophrenia was examined using Item Response Theory (IRT). Baseline PANSS assessments were analyzed from two multi-center clinical trials of antipsychotic medication in chronic schizophrenia ($n=1872$). Generally, the results showed that the PANSS (a) item ratings discriminated symptom severity best for the negative symptoms; (b) has an excess of "Severe" and "Extremely severe" rating options; and (c) assessments are more reliable at medium than very low or high levels of symptom severity. Analysis also showed that the detection of statistically and non-statistically significant differences in treatment were highly similar for the original and IRT-modified PANSS. In clinical trials of chronic schizophrenia, the PANSS appears to require the following modifications: fewer rating options, adjustment of 'Lack of judgment and insight', and improved severe symptom assessment.

1. Introduction

1.1. Background to the psychometrics of the Positive and Negative Syndrome Scale

The Positive and Negative Syndrome Scale (PANSS) aims to provide a comprehensive measure of the severity of symptoms of schizophrenia (Kay et al., 1987). It consists of 18 items from the Brief Psychiatric Rating Scale (Overall and Gorham, 1962) and 12 items from the Psychopathology Rating Schedule (Singh and Kay, 1975). Thus the PANSS consists of 7 items measuring positive symptoms, 7 items measuring negative symptoms, and 16 items measuring general psychopathology. It is widely used in clinical trials of schizophrenia to assess symptom severity, treatment response (Strous et al., 2003; Corrigan et al., 2004; Lieberman et al., 2005; Schooler et al., 2005; Leucht et al., 2007; Kinon et al., 2009) and remission (Andreasen et al., 2005; Leucht et al., 2009). Generally, research examines the validity and reliability of the PANSS based on classical psychometric theory. Classical psychometric theory, however, focuses on the aggregate scale or symptom cluster levels of reliability and validity, and not at the item or symptom level as examined by Item Response Theory (IRT).

Past validity research has empirically identified PANSS items that covary, known as factors. Generally, but not always (e.g., van der Gaag et al., 2006), factor analyses of the PANSS produce a five-factor solution (Bell et al., 1994; Lindenmayer et al., 1995a; Lindenmayer et al., 1995b; Marder et al., 1997; Higashima et al., 1998; Lancon et al., 1998; Lancon et al., 2000; Lykouras et al., 2000; Wolthaus et al., 2000; Emsley et al., 2001; Emsley et al., 2003; Levine and Rabinowitz, 2007). In a review of 20 factor analytic PANSS studies, 10 of the 30 items (Grandiosity, Stereotyped thinking, Somatic concern, Tension, Mannerism/posturing, Disorientation, Lack of judgment/insight, Disturbance of volition, Preoccupation, and Active social avoidance) did not consistently relate to one of the key 5 PANSS factors, suggesting that it is appropriate to modify the PANSS (Lehoux et al., 2009).

Another psychometric property of the PANSS that has been examined is reliability, referring to the accuracy of measurement. Reliability is generally measured with the reliability coefficient alpha (e.g., Cronbach's alpha) where values over 0.9 are very good, values of 0.8 to 0.9 good, values of 0.7 to 0.8 satisfactory, and values under 0.7 are not suitable for diagnostic use, since they contain an excess of error (Nunnally and Bernstein, 1978; Gorsuch, 1993; Kline, 1993). For instance, reliability of 0.70 means that 70% of the measure reflects the 'true score' (i.e., symptom severity), while the remaining 30% is attributable to measurement error. This source of error is consequential when predictive methods are used to examine efficacy (e.g., regression or survival modeling to predict outcomes). For instance, research indicates that the Negative factor (e.g., Emotional withdrawal) has acceptable reliability (0.89) in early onset patients (Emsley et al., 2003). The Anxiety and Depression factor, consisting of, for example, Tension ratings, has been reported to have low reliability (0.66) in early onset patients (Rabinowitz and Davidson, 2001). Similarly two studies report reliability of under 0.7 for the PANSS Depression factor (e.g., Lindenmayer et al., 1994; Mass et al., 2000). Accordingly, validity and reliability studies of the PANSS based on classic psychometric theory suggest that it is appropriate to improve the PANSS. To improve the reliability of measures, such as PANSS, modern

* Corresponding author. Tel.: +972 524896083; fax: +972 3 7617374.
E-mail address: levins@biu.ac.il (S.Z. Levine).

0165-1781/$ – see front matter © 2011 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.psychres.2011.03.014

psychometric theory offers a unique approach known as Item Response Theory (Embretson and Reise, 2000; Baker, 2001; Emsley et al., 2003; Reise and Henson, 2003).

1.2. Item Response Theory

IRT extends the classical psychometric view of reliability. The key terms of IRT and examples of its relevance to the PANSS are summarized in Table 1. In classical psychometric test theory reliability estimates are an aggregate value (e.g., Cronbach's alpha of the PANSS negative factor). In contrast, IRT permits the disaggregation of reliability by symptom severity level. For instance, in contrast to classic views of reliability, IRT acknowledges the possibility that the PANSS is reliable at moderate and not high or low Negative symptom severity levels. This is known as information accuracy (see Table 1).

IRT acknowledges that psychological assessments and measures, like the PANSS are inaccurate and probabilistic, unlike non-probabilistic concrete measures (e.g., meters). This may be contrasted to a PANSS assessment of hallucinations. For instance, the same patient could be rated as "Severe" and "Extremely severely" at the same time by different raters. Even with successful training 20% error has been reported between PANSS raters (Muller and Wetzel, 1998). To account for such issues, IRT permits examination of the extent to which observed ratings accurately discriminate latent symptom severity levels. In particular, IRT acknowledges that there is a discrepancy between the symptom severity ratings assigned and the existing (latent) symptom severity. Thus, the severity ratings that are rated do not entirely capture the latent symptom severity trait (see Table 1). Regarding IRT analysis of the PANSS, the ratings observed are the independent variables of symptom severity, whereas symptom severity is under observation, latent and is the dependent variable. This reflects the task of rating that requires the rater to estimate or predict latent symptom severity. This means that using IRT it is possible to estimate the extent that each rating option (e.g., 'Very severe') corresponds with the magnitude of a person's latent symptom severity on a specific item (e.g., Hallucinations). For example, a "severe" rating of hallucinations may under-estimate the true severity of the symptom, due to the difficulty associated with a patient who is in an exacerbated state.

IRT has already been used to examine and improve various psychiatric measures (e.g., Thompson et al., 2000; Young et al., 2003; Bagby et al., 2004; Minor et al., 2004; Rush et al., 2006; Bernstein et al., 2007; Schultz-Larsen et al., 2007). To date only one study has used IRT to examine the properties of the PANSS (Santor et al., 2007). Baseline PANSS scores were examined from 9205 patients with schizophrenia or a schizoaffective disorder from a large, naturalistic, observational study and from 12 randomized, double-blind, clinical trials comparing olanzapine to other antipsychotic drugs at various stages of illness. The study concludes by suggesting that most PANSS items perform well, although improvement or revisions to items and response options are recommended. Specific items that were reported to perform poorly include: Hostility, Difficulty in abstract thinking, Stereotyped thinking, Somatic concern, Mannerisms and posturing, Depression, Unusual thought content, Disorientation poor attention, Lack of judgment and insight, Disturbance of volition and Preoccupation. It was concluded that Positive and Negative items may constitute a "mini PANSS" that may be more reliable, require shorter administration and training time, and possibly reduce sample sizes needed for future research (Santor et al., 2007). Other research (Obermeier et al., 2010), using different analytic techniques, has suggested that the PANSS response options should be rescaled from 1 to 7 to 0 to 6 as to reduce inaccuracies in computing percent change and when applying significance testing (Leucht et al., 2010). This replicates early research advocating that the BPRS, a subset of PANSS items, be rescaled from 0 to 6 (Thompson et al., 1994).

1.3. Summary

Recent reviews (Lehoux et al., 2009; Leucht et al., 2010) and research (Thompson et al., 1994; Santor et al., 2007; Obermeier et al., 2010) suggest that despite its established status, the PANSS requires modification. IRT extends traditional psychometrics, and so may help improve the PANSS. First, IRT may identify superfluous items, thereby making possible the dropping of certain items while improving scale reliability and reducing administration time. Second, it is possible to examine the extent to which response options accurately discriminate symptom severity levels. Third, IRT emphasizes fine-graded symptom reliability (e.g., Hallucinations etc.) and to a lesser extent symptom groupings, whether by factors or subscales (e.g., Positive etc.). This means that IRT focuses on how each distinct item (i.e., symptom) operates rather than estimating the reliability of a scale (syndrome). Fourth, past IRT PANSS research has not specifically focused on chronic schizophrenia (Santor et al., 2007). The current study uses data from two large clinical trials of antipsychotic medication in chronic schizophrenia, and aims to examine and quantify the extent to which the PANSS: rating options accurately reflect symptom severity, items are adequate, and provides reliable information at varied symptom severity levels.

Table 1
Correspondence between standard Item Response Theory terminology and the PANSS.

Term	Definition	PANSS relevance
Key terms		
Classical test theory	Traditional psychometric methods (e.g., Cronbach's α)	Assesses syndromes not symptoms
Polychotomous responses	Measures with more than two response options These are observed responses.	E.g., "Severe", "Extreme"
Item	A question in a measure	The 30 PANSS questions E.g., Hallucinations.
Theta **θ**	Latent construct assessed with a scale	Symptom severity is latent, whereas PANSS ratings are observed.
Assessment method: parameter estimates		
Discrimination parameter (**α**)	Index of the strength the relationship between item and the latent construct it aims to measure.	Acknowledges the disparity between the hallucinations item assessment and latent hallucination severity
Threshold parameter (**ß**) — difficulty	Parameter to index the severity of an item response along the latent (**θ**) continuum of item response categories	Acknowledges the disparity between the severity rating (e.g., "Severe") and latent symptom severity
Assessment method: pictured graphically		
Information curves	Indexes items or scales values over latent severity (**θ**) to assess reliability at different severity levels	E.g., PANSS items or scales *may* be less reliable if latent symptom severity is high.
Item characteristic curve	Probabilistic index of the relationship between ratings on each category for an item and level on the latent construct (**θ**).	E.g. examines the distinction between "severity" and "Extreme severity" according to latent severity.