Correlations between self- and observer-ratings of psychopathology in at-risk mental state and first episode psychosis patients - influence of disease stage and gender

Short running title: Rater agreement in emerging psychosis

Andrea Spitz¹, Erich Studerus¹, Susan Koranyi², Charlotte Rapp³, Avinash Ramyead¹, Sarah Ittig¹, Ulrike Heitz¹, Martina Uttinger¹, Anita Riecher-Rössler*

¹ University of Basel Psychiatric Clinics, Center for Gender Research and Early Detection, Basel, Switzerland
² Jena University Hospital, Friedrich Schiller University Jena, Institute of Psychosocial Medicine and Psychotherapy, Jena, Germany
³ Psychiatric Services, Treatment Center for Psychosis, Solothurn, Switzerland

*Corresponding author:
Prof. Anita Riecher-Rössler, MD
Psychiatric University Clinics Basel
Center for Gender Research and Early Detection
Kornhausgasse 7
CH-4051 Basel, Switzerland
Phone: +41 61 325 81 61
Fax: +41 61 325 81 60
E-mail: anita.riecher@upkbs.ch

Word count main text: 2801
Word count abstract: 250
Abstract

**Aim:** Research findings on the correlations between self- and observer-ratings of schizophrenic psychopathology are inconsistent and have rarely considered first episode psychosis (FEP) and at-risk mental state (ARMS) for psychosis patients. This study investigates these correlations in ARMS and FEP patients and how they are moderated by disease stage and gender.

**Methods:** In the Basel Früherkennung von Psychosen (FePsy) study, positive and negative psychotic and affective symptoms were rated in 126 ARMS and 94 FEP patients using two observer and three self-rating scales. The agreement between self- and observer-ratings and the moderating influence of disease stage and gender was quantified using Pearson correlation and multiple regression models.

**Results:** Correlations between self- and observer-rated subscales covering the same symptom dimension were low and mostly non-significant except for one correlation of positive and one of negative symptoms. There was no moderating influence of disease stage, and gender on the correlations between self- and observer-ratings except for one higher association in positive symptoms in FEP compared to ARMS and in women compared to men. However, these significant interaction effects did not withstand correction for multiple testing.

**Conclusions:** This study suggest that the agreement between self- and observer-ratings in FEP and ARMS patients is rather low, similar across symptom dimensions, and only partially dependent of disease stage and gender. However, low correlations between self- and observer-ratings do not necessarily indicate that these patients have difficulties reporting their symptoms. They could also have occurred because the scales did not exactly cover the same symptom dimensions.

**Keywords:** gender, observer-rating, psychosis, self-rating
1. Introduction

The coherence between self- and observer-ratings assessing psychopathology in schizophrenic psychoses is an emerging topic in current research.\(^1\)\(^-\)\(^4\) Observer-rating scales are widely used\(^1\)\(^,\)\(^4\)\(^,\)\(^5\) and considered objective measures of the severity of psychopathological symptoms in patients with a psychotic disorder\(^6\), but they need a well-trained professional and are time-consuming.\(^1\)\(^,\)\(^3\) Self-rating scales, on the other hand, are more easily applicable. However, it is unclear whether self- and observer-rating scales measure similar constructs and whether psychosis patients are able to report their symptoms with sufficient accuracy.\(^7\)

Since schizophrenia patients have many features (e.g. poor insight, denial, delusions, cognitive deficits) that could hinder an accurate self-rating of their symptoms, it has long been assumed that self-ratings – especially of positive psychotic symptoms – are unreliable in these patients.\(^8\)\(^,\)\(^9\) However, our literature research revealed that at least six studies found a good agreement between self- and observer-ratings of positive psychotic symptoms in psychosis patients\(^3\)\(^,\)\(^4\)\(^,\)\(^9\)\(^-\)\(^12\) and only three studies found poor correlations.\(^13\)\(^-\)\(^15\)

With respect to negative symptoms, the concordance between self- and observer-rating scales seems to be rather inconsistent. Some studies found negative symptoms to be more difficult to be accurately reported than positive symptoms,\(^3\)\(^,\)\(^9\)\(^,\)\(^10\) whereas other studies suggested that even patients with a schizophrenic, schizoaffective or acute psychotic disorder are able to accurately report them.\(^11\)\(^,\)\(^12\) A further study by Bottlender et al.\(^16\) found equivocal results as it showed good agreement in the SANS total score but not in the subscales Apathy, Alogia, and Attention.

Studies assessing the concordance between self- and observer-ratings of depressive symptomatology in psychosis patients mostly showed good agreements.\(^13\)\(^,\)\(^15\)\(^,\)\(^17\) A recent study identified 49.2% of the patients to have equal self- and observer-ratings in depressive symptoms.\(^1\) However, Lasalvia et al.\(^14\) found significant correlations between affective symptoms only in non-psychosis but not in psychosis patients.
One explanation for these inconsistent results is that in many studies self- and observer-rating scales did not tap exactly the same symptom dimension. Additionally, existing studies vary in several factors that can potentially moderate the relationship between self- and observer-ratings, such as disease state, diagnostic group, degree of insight, and gender distribution. However, only few studies have investigated the influence of these moderating factors. Below, we will summarize the literature regarding the influence of disease stage, and gender, as our study will focus specifically on these factors.

To our knowledge, no study has investigated whether the agreement between self- and observer-ratings differs between patients with a first episode psychosis (FEP) and those that have an at-risk mental state (ARMS) for psychosis. Existing studies have only focused on one of these disease stages. FEP patients were found to have a good association in positive, but not in negative symptoms,\textsuperscript{10} whereas ARMS patients were shown to have more psychosis risk symptoms in self-reports than in clinical interviews.\textsuperscript{18} However, since lack of insight can lead to a decreased agreement between self- and observer-ratings\textsuperscript{4} and since FEP patients were found to have more impaired insight than ARMS patients,\textsuperscript{19} it is reasonable to assume that a direct comparison between ARMS and FEP patients would show a higher agreement in ARMS patients.

Although gender differences in schizophrenia have been described in almost all aspects, including age of onset, incidence, symptomatology, treatment response, and outcome,\textsuperscript{20-22} little is known about whether gender influences the agreement between self- and observer-ratings. Some studies showed a higher agreement of affective symptoms in women compared to men in mixed patient samples or patients with depression.\textsuperscript{23, 24} However, other studies do not support these findings in samples with psychotic and a non-psychotic major depression.\textsuperscript{17, 25} To our knowledge, only one study investigated the influence of gender on the agreement between self- and observer-ratings of positive psychotic symptoms in schizophrenia patients.\textsuperscript{4} This study found no influence of gender.
To improve on previous studies, this study aimed to compare self- and observer-ratings of affective, negative and positive symptoms in both ARMS and FEP patients. Furthermore, we aimed to investigate whether the agreement was dependent on disease stage and gender. We hypothesized that the association between self- and observer-rating is higher in ARMS than in FEP patients and higher in women than in men.

2. Methods

2.1 Setting and Recruitment

All data were collected as part of the Basel Früherkennung von Psychosen (FePsy) project, a prospective multilevel study, which aims to improve the early detection of psychosis.26, 27 The study was approved by the ethics committee of the University of Basel, and all participants provided written informed consent. Patients were recruited from the 1st of March 2000 to the 31st of January, 2013 via the FePsy Clinic, which was specifically set up to identify, assess, and treat individuals in the early stages of psychosis.

2.2 Screening procedure

Screening of ARMS and FEP patients was performed with the Basel Screening Instrument for Psychosis (BSIP), which has been shown to have a good interrater reliability (K=0.67) and a high predictive validity.28 Individuals were classified by the BSIP as being in an ARMS for psychosis, having a FEP, or being not at risk for psychosis using criteria corresponding to those of Yung et al.29

2.3 Assessment of psychopathology

The Brief Psychiatric Rating Scale Expanded Version (BPRS-E)30 and the Scale for the Assessment of Negative Symptoms (SANS)31 were used as observer-ratings and the
Frankfurt Complaint Questionnaire (FCQ)\textsuperscript{32}, the Selfscreen-Prodrome (SSP)\textsuperscript{33} and the Paranoid Scale (PS)\textsuperscript{34} were used as self-ratings of psychopathological symptoms.

The BPRS-E is a widely used rating scale for assessing general psychopathology and consists of 24 items, which can be grouped to the four subscales Depression/Anxiety, Psychosis, Negative Symptoms and Activation.\textsuperscript{35} All BPRS-E items are rated on a 7-point severity scale.

The SANS is a 24-item scale for assessing negative symptoms. The items of the SANS are rated on a five-point ordinal scale and are grouped to 5 subscales: Affective Flattening, Alogia, Apathy/Avolition, Anhedonia/Asociality and Attention.

The FCQ contains 98 dichotomous items and is used to assess so-called “basic symptoms”, which are abnormal subjective experiences that can occur in a prodromal state of psychosis and that seem to have a predictive validity for the onset of psychosis.\textsuperscript{36} These symptoms have been called “basic” to indicate their proximity to hypothesized basic neural dysfunctions of schizophrenia.\textsuperscript{37} The FCQ contains four factors: Depression, Disturbances of automated responses, Perceptual disturbances and Overinclusion.\textsuperscript{32}

The PS consists of a subset of items of the Paranoid Depression Scale (PDS) that contains paranoid and depressive symptoms. The PS comprises 14 items which can be grouped into the three subscales Paranoid Tendencies, Test Motivation and Denial of Illness.

The SSP is a screening instrument to identify patients with a risk for psychosis. It consists of 32 dichotomous items regarding prodromal and pre-psychotic symptomatology.

All observer measures were conducted by well-trained psychologists or psychiatrists.

2.3 Statistical analysis

Analyses were performed using the Statistical Package for the Social Sciences (SPSS) for Windows, version 19, and the R environment for statistical computing.\textsuperscript{38} Differences in
socio-demographic and clinical characteristics between ARMS and FEP patients were tested with $t$ and $\chi^2$ tests.

First we used Pearson correlations to compare the already existing and published subscales of the five self- and observer-rating instruments. However, because these rating scales frequently differ in their item content, we also constructed new subscales from self-rating items that were as similar as possible with original scales in the BPRS and SANS. Specifically, by applying hierarchical item cluster analysis and based on theoretical knowledge about the dimensional structure of psychopathology, we grouped the items of each self-rating scale to the subscales Affective Symptoms, Positive Symptoms and Negative Symptoms in such a way that they were most similar to BPRS Depression/Anxiety, BPRS Psychosis, and SANS total score, respectively. For assessing negative symptoms, we used the SANS total score instead of the BPRS subscale for negative symptoms because it is covering this symptom dimension more completely and reliably. With the PS items, only the new subscale (“Positive Symptoms”) was formed. The items of each newly formed subscale are shown in Supplementary Table 1. To evaluate the internal consistency and homogeneity of the new subscales, Cronbach’s $\alpha$ and Revelle’s $\beta$ were calculated. In case of dichotomous and polytomous items, these measures were based on tetrachoric and polychoric correlations, respectively.

To evaluate the correlations between all self- and observer-rated scales we generated a Multitrait-multimethod matrix (MTMM). This approach evaluates the construct validity of measures of different concepts assessed by different methods. It shows how the correlations between different measures vary as a function of different item content and method.

Although all psychopathological assessments were obtained at baseline, they were not always obtained at the same visit. Hence, in accordance with previous studies, we correlated only those measures of each patient that were obtained within a period of seven days.
Rater agreement in emerging psychosis

To examine whether gender and disease stage moderate the correlations between self- and observer-ratings, multiple regression models with the observer-rating scale as dependent variable, the self-rating scale as the first independent variable and disease stage or gender as the second independent variables and the interactions between these variables were performed. To facilitate interpretation, continuous variables were z-transformed.

To correct for multiple testing, p-values were adjusted using the Benjamini-Hochberg method. 42

3. Results

3.1 Sample description

Socio-demographic sample characteristics are presented in Table 1. ARMS did only differ from FEP patients regarding age (t(185) = -3.69, p < .001).

Clinical characteristics of our sample are shown in Table 2. FEP patients had higher scores in all scales assessing positive and basic symptoms. However, they did not differ in negative symptoms scales except for a higher score of FEP patients in the newly constructed self-rating SSP Negative Symptom scale. With regard to affective symptoms, FEP scored higher in the BPRS Depression/Anxiety and self-rating FCQ Affective Symptoms scale, but not in the self-rating SSP Affective Symptoms scale.

3.3 Associations between self- and observer-ratings

Correlations of the original subscales between self- and observer-ratings are illustrated in Figure 1. The highest correlations between subscales with similar item content were between FCQ Perceptual disturbances and BPRS Psychosis (r(63) = .342, p adjusted = .056)
Rater agreement in emerging psychosis

as well as between PS Paranoid Tendencies and BPRS Psychosis \((r(70) = .455, p \text{ adjusted} = .002)\).

Correlations between the self- and observer-ratings, internal consistencies, homogeneities, and sample sizes of the newly constructed subscales are illustrated in Table 3. The newly constructed subscales showed a good internal consistency \((\alpha = .86 - .96)\) and homogeneity \((\beta = .7 - .85)\). However, internal consistencies of the BPRS Psychosis and Depression/Anxiety subscales were \(\alpha < .8\) and \(\alpha < .7\), respectively. Heterotrait-monomethod correlations were higher than monotrait-heteromethod correlations, suggesting that there was more common variance due to the method than the content. There were only two significant correlations between self- and observer-rating scales covering the same symptom dimension. Specifically, the FCQ Negative Symptoms subscale correlated significantly with the SANS scale \((r(66) = .317, p \text{ adjusted} = .021)\) and the PS Positive Symptoms correlated significantly with the BPRS Positive Symptoms subscale \((r(70) = .454, p \text{ adjusted} < .001)\).

3.4 Influence of disease stage and gender on the association between self- and observer-ratings

As shown in Figure 2, there were no Group × Self-rating scale interactions in affective and negative symptomatology. However, in positive symptoms, there was one significant Disease stage × Self-rating scale interaction with the FCQ Positive Symptom scale, \(R^2 = .534\), \(F(1,63) = 7.38, p = .009\), \(\eta^2 = .108\), which was due to a higher correlation between self- and observer-rating in FEP than in ARMS patients. However, when corrected for multiple testing, this interaction effect was only significant at a trend level \((p \text{ adjusted} = 0.060)\).
In the analyses including gender, there was only one statistically significant Gender × Self-rating scale interaction, namely, with the SSP Positive Symptoms subscale, $R^2 = .168$, $F(1,53) = 6.009$, $p = .018$, $\eta^2 = .105$, suggesting that women showed a higher correlation of this subscale with the BPRS Positive Symptom scale than men (Figure 3). However, this significant interaction effect did not withstand correction for multiple testing ($p$ adjusted = 0.124).

Insert Figure 3 about here

4. Discussion

The aim of this study was to investigate the correlations of self and observer-ratings in ARMS and FEP patients and the influence of disease stage and gender on these correlations. Using the original subscales, we found relatively high correlations in positive symptom dimensions but not in the other symptom dimensions. When the scales were adapted to have better matching item contents, we found two significant correlations between self- and observer-ratings covering the same symptom dimension, namely, one with positive and one with negative symptoms. Furthermore, disease stage and gender each moderated one pair of self-observer-ratings, but only if $p$-values were uncorrected for multiple testing.

The construction of new subscales with more homogeneous and better matching item content improved the interpretability of the results and led to higher self-observer agreements, particularly in the domain of negative symptoms. However, overall the correlations were still relatively small. Although all seven pairs of subscales covering the same symptom dimension correlated positively, only two were statistically significant, indicating that the agreements between self- and observer-ratings were rather low. Furthermore, no clear pattern emerged with regard to strength of association and symptom dimension. Since we found statistically significant correlations with both positive and negative symptoms, our results do not confirm earlier findings of Hamera et al.\textsuperscript{9} and Preston
and Harrison\textsuperscript{10} according to which negative symptoms are more difficult to be accurately reported than positive symptoms but support earlier findings of Bell et al.\textsuperscript{12} and Liraud et al.\textsuperscript{11} The lack of association between scales measuring affective symptoms stands in contrast to previous studies which reported good agreements in this dimension.\textsuperscript{1, 15}

With regard to the moderating influence of the disease stage, we could not confirm that the association between self- and observer-rating is higher in ARMS than in FEP patients. However, when \( p \)-values were not corrected for multiple testing, there was one significant interaction which was in the opposite direction of what we had expected. Specifically, FEP showed higher correlations than ARMS patients between the BPRS and FCQ Positive Symptom scales. One possible explanation is that lower occurrence of positive psychotic symptoms in the ARMS group led to a distribution of positive symptoms with a lower spread and higher positive skew than in the FEP group which in turn might have led to a stronger attenuation of the correlation.

With regard to the moderating influence of gender, we found that women showed a higher association between BPRS and SSP Positive Symptoms than men, suggesting that women are more accurate in reporting their positive psychotic symptoms. This finding stands in contrast to the study of Lincoln et al.\textsuperscript{4} which did not find an influence of gender on the rating of positive psychotic symptoms. However, the gender effect we found should be interpreted with caution because there was no influence of gender in the two other comparisons regarding positive symptoms (i.e. BPRS vs. FCQ Positive Symptoms and BPRS vs. PS Positive Symptoms) and the effect was only significant when \( p \)-values were not corrected for multiple testing. Furthermore, our results did not support earlier studies demonstrating that women report their affective symptoms more accurately than men.\textsuperscript{23, 24} However, these studies are difficult to compare with our study because they were based on mixed patient samples.
Rater agreement in emerging psychosis

Our study has some limitations. Firstly, even though we had improved the comparability of the scales by forming new subscales, we were quite limited in the item content and thus it is possible that our subscales still insufficiently covered the same symptom dimensions. Other studies solved this problem using newly constructed self-rating questionnaires, modified observer-ratings to self-questionnaires or concentrated their analysis on a special symptom dimension. Secondly, although we had obtained self- and observer-ratings from 220 patients in total, a relatively large proportion of these had to be excluded because the time difference between self- and observer-rating was too large. Thirdly, we assumed according to the literature that observer-ratings are closer to “the truth” and therefore represent the gold standard. However, in order to verify this we would have to link both types of assessment to an external criterion. Future studies should directly assess the value of both types of assessment. For example, it is likely that observer ratings are better for an accurate diagnosis whereas self-ratings provide additional information that can help to increase treatment compliance.

Taken together, we found that the associations between self- and observer-ratings were rather low. Contrary to our expectations, they were neither higher in ARMS than in FEP patients, nor higher in women than in men when corrected for multiple testing. The results of our study therefore imply that self-rating scales cannot be a substitute for the more time-consuming observer-rating scales and vice versa.

Acknowledgment

This work was supported by the Swiss National Science Foundation (grant numbers 3200-057216.99, 3200-0572216.99, PBBSB-106936, and 3232BO-119382); the Nora van Meeuwen-Haefliger Stiftung, Basel (CH).

All authors do not have any conflict interests that might be interpreted as influencing its content.
References


Rater agreement in emerging psychosis


Rater agreement in emerging psychosis


