Site vs. Remote Inter-Rater Reliability of the PANSS and Information Demand

Bethanne Friedmann, PsyD[1], Henry Riordan, PhD[1], Erin Kornsey, MS[1], Evan Braxton[1], Michael F. Murphy, MD, PhD[1], Neal R. Cutler, MD[2]

[1]Worldwide Clinical Trials, King of Prussia, PA, [2]Worldwide Clinical Trials, Beverly Hills, CA

Abstract

Psychiatric clinical trials utilize subjective measures to assess efficacy. Training and alignment of raters is crucial to ensure adequate inter-rater reliability and eventual study success. Most training is conducted at study start-up and focus on raters' ability to score video recordings within an acceptable standard. While this method is proven to align raters, it does not assess a rater's ability to conduct an interview, nor does it prevent rater drift during the study. There are numerous monitoring methods that assess interview skills as well as decrease rater drift.

Videotaping/videoconferencing promote continuous inter-rater reliability by permitting expert independent ratings of the same subject and immediate feedback on scoring and interview skills. Accurate scoring of the Positive and Negative Syndrome Scale (PANSS) requires rater observation, interviewing the patient (clinical interview) and collecting information from a family member over the previous week (informant information). However, remote raters may not have access to all of the information needed to assess all scale items resulting in poor validity and reliability.

For this study, six raters completed and videotaped twenty-eight ratings of the Structured Clinical Interview for the Positive and Negative Syndrome Scale (SCI-PANSS). These videos were rated by expert same-language raters. Results demonstrated only a moderate agreement between the site raters and expert raters for total PANSS scores. PANSS items were divided into three categories based on the type of information needed to score each item. Average Intraclass Correlation Coefficient (ICC) suggested only poor reliability between site raters and expert raters for items based solely on informant information (ICC =.04); fair reliability was found for items scored using the clinical interview and informant information (ICC = .376). Results indicate there was a moderate reliability for items scored solely from the clinical interview (ICC = .403). This data underscores the demand for all available source of information for remote raters in order to ensure valid and reliable PANSS assessments.

Background

The PANSS[1] is a 30-item, 7-point interview-based assessment utilized to measure symptoms of psychosis in a variety of psychiatric disorders, such as schizophrenia, bipolar disorder, and depressive psychosis. It is routinely used in psychopharmacological studies to measure change in psychotic symptoms. It is divided into seven positive (P1-P7), seven negative (N1-N7), and 16 general items (G1-G16) associated with the symptoms of psychosis.

Utilization of the structured clinical interview for the PANSS (SCI-PANSS) increases the validity of the information obtained. The worst symptoms from the past seven days were assessed. Scoring is based on information obtained during an interview with the patient and corroborated by an informant. Collateral information can come from primary care hospital staff and family members, in addition to behavioral observations by the rater during the clinical interview.

To ensure the validity of data, site raters (SR) were monitored by expert raters (ER) via video recorded interviews. We hypothesized that lack of access to the informant would differentially affect PANSS items based on the amount of collateral information needed for accurate rating.

Methods

PANSS items were divided into three categories according to the type of information needed to score these items appropriately (see Table 1). One category that has two items (N4 and G16) contains items based solely on informant report (I); the second category has 16 items (P2, N1, N3, N5-7, G1-4, G9-13 and G15) that utilize information gathered during the clinical interview (C); the third category is composed of the 12 remaining items that utilize information from both informant and clinical interview (IC).

CATEGORY	ITEMS
Informant Information Items (I)	N4, G16
Clinical Interview Items (C)	P1, P3, P4, P5, P6, P7, N2, G5, G6, G7,G8, G14
Both Informant Information and Clinical Interview (IC)	P2, N1,N3, N5, N6, N7, G1, G2, G3, G4, G9, G10, G11, G12, G13, G1

Table 1. PANSS categories divided by information sources.

This study used six raters from a multinational Phase IIa clinical trial investigating the effects of an antipsychotic agent on individuals diagnosed with schizophrenia who were hospitalized for an acute exacerbation of psychotic symptoms. The primary efficacy of the clinical trial was a change in PANSS score from baseline to week four.

The three expert raters used in this study were considered key opinion leaders in schizophrenia in Russia and Ukraine. They all held medical degrees and university appointments.

Table 2 depicts the experience of the site raters. All six raters were psychiatrists. Three raters had 3-5 years experience administering the PANSS, and three raters had more than 6 years experience administering the PANSS. All raters had used the PANSS multiple times in the past two years (three raters administered the PANSS 26-50 times; three raters administered the PANSS more than 50 times.)

NUMBER OF TRIALS CONDUCTED IN THE PAST TWO YEARS Mean (Standard Deviation)	RESEARCH EXPERIENCE (YEARS) Mean (Standard Deviation)	CLINICAL EXPERIENCE (YEARS) Mean (Standard Deviation)
3.4 (1.14)	5.4 (1.67)	7.8 (5.54)

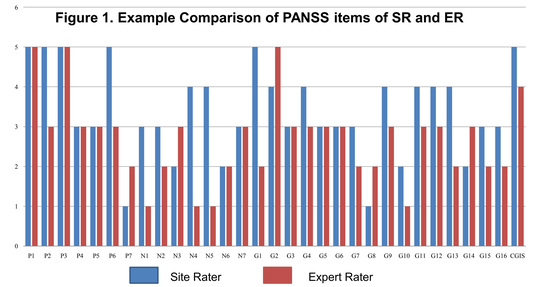
Table 2. Site Raters' Experience with Schizophrenia.

Methods (cont)

The site raters were trained on how to conduct the PANSS at an investigators meeting in Europe where they participated in an interactive review of anchors and scoring conventions which was simultaneously translated into their native language. Certification was based on raters' ability to score a video-recorded patient interview. The video interview was conducted in English and subtitled in their native language. Site and expert raters were aligned to a consensus scoring of the certification videos and had to achieve greater than or equal to 80 percent concordance prior to being approved for this trial.

Twenty-eight video recordings of site raters PANSS interviews were obtained from visits 3 and 9. All site raters utilized the SCI-PANSS. Each PANSS interview was conducted in a local language and rated separately by an in-country expert rater. All site raters were instructed to provide a summary of corroborating information. Expert raters were blinded to the site rater's scores and did not have direct access to informants; all information had to come from the videotape, including site rater narratives of informant information. The expert raters provided feedback on the interview with their scores for comparison to the site raters' scores. Figure 1 demonstrates an example comparison between a site rater and expert rater by individual PANSS items.

Figure 1. Example Comparison of PANSS items of SR and ER



The ICC was utilized as a measure of the reliability of ratings between site and expert raters and may be conceptualized as the ratio of between-groups variance to total variance.

The ICC evaluates the level of agreement between raters in measurements, when the measurements are interval in nature. This method is better than ordinary correlation as more than two raters can be included, and there is a correction for correlations between raters that becomes apparent when the range of measurement is large. The coefficient represents concordance, where 1 is perfect agreement and 0 is no agreement at all.

Results

ICC by category are shown in Table 3. Average ICCs suggested poor reliability between site raters and experts for items based on informant report only (ICC I =.04); fair reliability (ICC IC=.376) for combined items, and moderate reliability (ICC C = .403) for items gathered based solely upon clinical interview. Despite these qualitative differences there were no significant statistical differences between these three information categories (x^2=1.68, p=.43) when comparing ICCs.

CATEGORY	INTRACLASS CORRELATION (ICC)
Informant Information Items (I)	.04
Clinical Interview Items (C)	.403
Both Informant Information and Clinical Interview (IC)	.376

Table 3. Intraclass Correlations (ICC) by Category.

According to interpretation conventions by Fleiss (1981)[2], the ICC for the overall PANSS score showed only moderate agreement between site and expert raters (ICC = .438, p >.05). The majority of individual PANSS items ICCs (60%) fell in the moderate range (.40-.59) of reliability while 20% of the site rater versus expert rater ICCs fell in the good to excellent range (0.60–1) of reliability, and 20% fell in the poor range (<0.20). No readily apparent pattern of ICCs was noted based on positive/negative/general psychopathology composites. Rater demographics such as prior PANSS experience and rating performance at the investigator meeting did not appear to be related to site versus expert ICCs.

Conclusions

The PANSS assessment is the cornerstone of efficacy analysis for a majority of clinical trials. Raters play a significant role in obtaining this data. This is typically site raters, but the use of remote raters is increasing. A multitude of studies fail to demonstrate change in symptoms based on PANSS scoring. It is critical that sources of variability among raters is minimized in order to ensure that PANSS assessments are accurate.

Videotaping / videoconferencing promote continuous inter-rater reliability by permitting expert independent ratings of the same subject and immediate feedback on scoring and interview skills. However, remote raters may not have access to all of the information needed to assess all scale items, resulting in poor validity and reliability.

By limiting what PANSS information remote expert raters had access to, this study showed that incomplete information, whether from informants or from the clinical interview, altered inter-rater reliability. Not surprisingly, the lack of clinical interview information most reduced inter-rater reliability. But the fact that the majority of individual PANSS items ICCs fell only in the moderate range of reliability between site and remote raters, suggests there may be overall reliability issues with remote raters that could affect PANSS scoring. Reliability may be improved by using the informant questionnaire for the PANSS (IQ-PANSS) as opposed to site rater narratives. Further study of remote rater PANSS scoring variability is warranted.

References

Kay S. R., Opler L.A., Fiszbein A. 2006. Positive and Syndrome Scale (PANSS) Technical Manual. New York. Multi-Health Systems.

Fleiss J.L. 1981. Statistical Methods for Rates and Proportions. 2nd ed. Wiley, New York.