Methods for Determining Inter-rater Reliability of the PANSS: a review of the literature

Kia Crittenden, PhD, Christian Yavorsky, PhD, Felice Ockun, MSW, MS, Kristy Wolanski, MA, Kenneth A. Kobak, PhD

Background: Over the past few decades, there has been a growing increase in the response to placebo as measured by the Positive and Negative Syndrome Scale (PANSS), one of the most commonly administered outcome measures in antipsychotic medication trials. This has resulted in an increasing rate of failed trials. An examination of the factors associated with clinicians’ interviews that play a role in this phenomenon is warranted. Inter-rater reliability (IRR) is an important factor that can impact CNS trials, as poor reliability increases error variance, which reduces study power and increases the risk for type II errors. The method for evaluating reliability impacts the reliability figures, e.g., observation of tapes results in artifically higher IRR than independent interviews, as the former artifically reduces variance information. Thus, the methodology for determining IRR is a critical factor in interpreting a study’s results. This study systematically reviewed the literature on the methods used in determining reliability of the PANSS and the relationship between the method and level of reliability achieved.

Method: We searched PubMed using keywords “PANSS” and “reliability.” A forward search was also conducted of cited references in identified citations.

Results: A total of 34 studies were identified reporting on the reliability of the PANSS. These 34 studies varied in terms of sample version, and included the PANSS, SCID-PANSS, SCI-PANSS, Kiddie-PANSS, and foreign language versions. In roughly half the studies (n=18; 53%), the method for determining reliability was described. The intraclass correlation coefficient (ICC) is the most common test statistic used in the methodology. The most common test statistic utilized was the intraclass correlation coefficient (ICC) (n=20; 59% of studies), followed by kappa (n=4; 12%), Pearson correlation (n=2; 6%), and unspecified (n=8; 26%). Bartko and Carpenter (1976) suggest using ICC when there are two or more raters, the data uses a continuous scale and the observations are not independent. The intraclass correlation coefficient for the positive PANSS scale score ranged from 0.56 to 0.99 when measured via taped observation and from 0.72 to 0.99 when measured via joint interview. The negative scale score inter-rater reliability ranged from 0.27 to 0.70 for taped observation and from 0.65 to 0.92 for joint interviews. Inter-rater reliability on total PANSS score ranged from 0.66 to 0.71 for tape observation and from 0.92 to 0.99 for joint interviews.

Conclusion: Methodology for testing IRR is not often cited. Tape rating is the most common method reported followed by joint interview. Reliability on the PANSS using joint or tape review methods appears good, with tape review slightly lower than joint interviews. It is difficult to determine if these statistics are true representations of agreement, as these techniques may artificially inflate IRR measures. If IRR is inflated, then the potential for error is increased and study power diminished. A more rigorous way to evaluate IRR is to use independent interviews which would allow each rater to administer the instrument to the same patient at different times (although not without its own problems, e.g., recency and latency effects) and thus more accurately measure reliability between raters. More rigorous methodology for evaluating IRR enables a more accurate evaluation of rater reliability. Furthermore, using the ICC is a more appropriate index of inter-rater reliability due to the lack of independence of the observations, less than half of the methodology. The methodology for determining IRR is an important interpretation of ICC, and should be described when reporting study results.