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 clinician 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 artificially higher IRR than independent interviews, as the former artificially reduces information variance. 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.

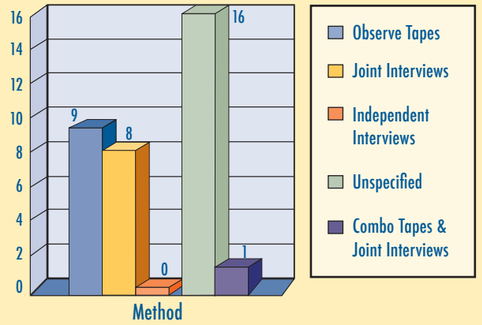
Method: We searched PubMed using keywords "PANSS" and "reliability." A forward search was also conducted of cited references in identified citations.

Results: A total of 34 studies were identified reporting on the reliability of the PANSS. These 34 studies varied in terms of PANSS 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. These included rating video or audio tapes (n=9; 26%), joint interviews, where one rater observes another in real time (n=8; 24%), and a combination of tape review and joint interviews (n=1; 3%). The rest of the studies (n=16; 47%) did not describe the methodology for determining reliability. No studies were identified where independent interviews were used to evaluate inter-rater reliability. In 19 of the 34 studies (56%) some type of rater training was mentioned as having been conducted prior to testing as part of 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 interviews. The negative scale score inter-rater reliability ranged from 0.27 to 0.90 for taped observation and from 0.63 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 utilize 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, though the ICC is a more appropriate index of inter rater reliably due to the lack of independence of the observations, less than half of the reviewed studies used this measure. Method of evaluating IRR is important in interpreting ICC, and should be described when reporting study results.

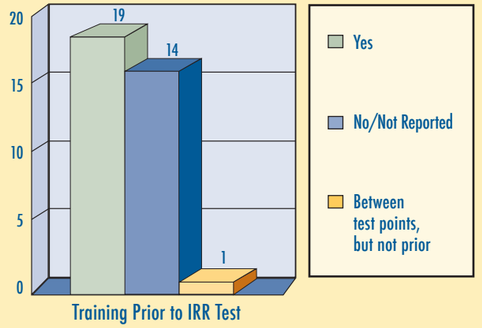
Bartko, J. J. & Carpenter, W. T. (1976). On the methods and theory of reliability. *The Journal of Nervous and Mental Disease*, 163, 307-317.

METHOD OF TESTING IRR FOR ALL VERSIONS



N=34

FREQUENCY OF RATER TRAINING CONDUCTED PRIOR TO TESTING IRR



IRR ACROSS REVIEWED STUDIES

	Instrument Version	Method of Testing IRR	Test Statistic	Positive Scale	Negative Scale	General Scale	Total Scale	Unspecified or Other
Resnick et al., (2003)	PANSS	Joint	1	0.99			0.99	0.93
Kay et al., (1991)	SCID-PANSS	Joint	1	0.96	0.92	0.92	0.92	
Peralta & Cuesta, (1994)	PANSS	Joint	1	0.72	0.80	0.56		
Norman et al., (1996)	PANSS	Joint	1	0.79	0.63	0.82		
Igarashi et al., (1998)	PANSS (Japanese)	Joint	1	0.85	0.83	0.75		
Kay et al., (1988)	PANSS	Joint	2	0.73-0.89	0.70-0.89	0.69-0.94		
Deutsch et al., (2003)	PANSS	Joint	3					0.83
Nakaya et al, (1997)	PANSS	Joint	3					0.65-0.95
Betsen et al., (1996)	PANSS (Norwegian)	Tape	1	0.90	0.74	0.89		
Garety et al., (2005)	PANSS	Tape	1	0.92-0.98				
Purine et al., (2000)	SCID-PANSS	Tape	1	0.79	0.77	0.36		
Lindstrom et al., (1994)	SCID-PANSS	Tape	1	0.98-0.99	0.83-0.90	0.95-0.98		
Muller et al., (1998)	PANSS	Tape	1	< 0.70	< 0.70	<0.70		
von Knorring & Lindstrom, (1995)	SCI-PANSS & PANSS	Tape	1	0.95-0.99 (SCI-PANSS) 0.56-0.77 (PANSS)	0.83-0.90 (SCI-PANSS) 0.27-0.43 (PANSS)	0.95-0.99 (SCI-PANSS) 0.56-0.77 (PANSS)		
von Knorring & Lindstrom, (1992)	PANSS (Swedish)	Tape	1	0.75-0.77	0.27-0.46	0.56-0.72	0.66-0.71	
Muller & Wetzel, (1988)	PANSS	Tape	3	0.69-0.91	0.43-0.80	0.50-0.90		
Muller & Davids, (1999)	PANSS	Tape	4	86%	93%	89%	89%	
Salyers et al., (2001)	PANSS	Tape & Joint	1	0.87	0.74	0.71		
Ritsner & Ratner, (2006)	PANSS	Unspecified	1					0.78-0.95
Fields et al., (1994)	Kiddie-PANSS	Unspecified	1	0.76	0.78	0.81	0.86	
Volavka et al., (2000)	PANSS	Unspecified	1				> or = 0.80	
Barch et al., (2004)	PANSS	Unspecified	1				0.97	
Bell et al., (1992)	PANSS	Unspecified	1	0.93	0.94	0.84	0.91	
Harris et al., (1997)	PANSS	Unspecified	1				> 0.70	
Gearon et al., (2004)	PANSS	Unspecified	1					0.63-0.93
Herz et al., (2000)	PANSS	Unspecified	2	0.94	0.80	0.89		
Troisi et al., (1997)	SCID-PANSS (Italian)	Unspecified	3				> or = 0.80	
Goff et al., (1999)	PANSS	Unspecified	5					> 0.80
Dixon et al., (2001)	PANSS	Unspecified	5					0.63-0.93
Goldberg et al., (2001)	PANSS	Unspecified	5					0.63-0.93
Erickson et al., (2005)	PANSS	Unspecified	5					> = 0.80
Pyne et al., (2003)	PANSS	Unspecified	5					0.92
Tait et al., (2005)	PANSS	Unspecified	5				within 80% of range	
Bouchard et al., (2000)	PANSS	Unspecified	5					minimum 80%

Test Statistic: 1 = intraclass coefficient, 2 = Pearson r, 3 = Kappa, 4 = other, 5 = unspecified

MEDAVANTE® RESEARCH INSTITUTE
Subjective —> Objective
MedAvante Inc., 100 American Metro Boulevard, Suite 106, Hamilton, NJ 08619