REVIEW

Positive and Negative Syndrome Scale (PANSS) Training: Challenges, Solutions, and Future Directions

by MARK G. A. OPLER, PhD, MPH; CHRISTIAN YAVORSKY, PhD; and DAVID G DANIEL, MD

Dr. Opler is Adjunct Assistant Professor at NYU School of Medicine and Chief Research Officer at MedAvante-ProPhase Inc. in New York, New York. Dr. Yavorsky is Chief Scientific Officer and Clinical Director at Cronos Clinical Consulting in Lambertville, New Jersey. Dr. Daniel is Chief Medical Director and Senior Vice President of Bracket Global, LLC and Clinical Professor at George Washington University, in Washington, DC.

Innov Clin Neurosci. 2017;14(11–12):77–81

ABSTRACT

ABSTRACT: Rater training and the maintenance of the consistency of ratings are critical to ensuring reliability of study measures and sensitivity to changes in the course of a clinical trial. The Positive and Negative Syndrome Scale (PANSS) has been widely used in clinical trials of schizophrenia and other disorders and is considered the "gold standard" for assessment of antipsychotic treatment efficacy. The various features associated with training and calibration of this scale are complex, reflecting the intricacy and heterogeneity of the disorders that the PANSS is used to evaluate. In this article, the authors review the methods for ensuring reliability of the PANSS as well as a proposed trajectory for its use in the future. An overview of the current principles, implementation, technologies, and strategies for the best use of the PANSS; tips for how to achieve consistency among raters; and optimal training practices of this instrument are presented.

KEYWORDS: Positive and Negative Syndrome Scale, PANSS, rater, rater training, technology, clinical trials

Rater training helps to ensure reliability in measurement throughout the course of a clinical trial. Precision in the use of a rating scale is important primarily because statistical power to detect differences between treatment groups increases proportionally to inter-rater reliability. A related secondary objective is to ensure that when scale items or subscale score thresholds are being incorporated as inclusion criteria, all raters in a study can reliably classify subjects. Rater training further enhances precision by standardizing interview procedures and codifying the principles of use for a given scale across raters, sites, and regions.[1]

The Positive and Negative Syndrome Scale (PANSS) has several complex features and requires a thorough and structured approach to rater training.[1] Compared with rating scales developed for other disorders, the PANSS has many items, evaluates a multidimensional array of symptoms (e.g. positive, negative, neuromotor, depressive), and involves the use of data from patient reports, caregiver reports, and clinical observations. Consequently, the PANSS takes up more time during training and requires a greater amount of time for one to master it compared to many other instruments.

As described in the original 1987 publication,[2] each PANSS item contains three elements that must be used correctly in order to ensure that reliability and validity are maintained:[1] 1) The item definition describes the construct under evaluation; 2) Each item contains a detailed description of the basis for rating that indicates the sources of information intended to be used for each item. These sources include observations made during the interview, the patient verbal report, and/or corroborative information obtained from caregivers about symptoms and behaviors during the reference period prior to the assessment; and 3) Each item includes a set of carefully written anchors for each level of severity, from 1 (absent) to 7 (extreme).

CORE PRINCIPLES IN THE USE OF THE PANSS

Several approaches to the use of the PANSS might help raters and those leading training programs to achieve a high degree of reliability. Four core principles, summarized here, are taken from publications and lectures given by Dr. Lewis Opler over the course of many years. We summarize them briefly here so as to provide guidance to individual raters

FUNDING: No funding was provided for this article.
DISCLOSURES: Dr. Opler is a full-time employee and shareholder of MedAvante-ProPhase Inc. Dr. Yavorsky is a full-time employee of Cronos CCS. Dr. Daniel is a full-time employee and shareholder of Bracket Global, LLC.
CORRESPONDENCE: Mark Opler PhD, MPH; Email: mgo4@caa.columbia.edu

and those persons implementing training programs to improve reliability.

First principle—Read each item definition and all anchor points carefully and interpret each element as literally as possible. The process of rating PANSS items requires a very close reading of each required element. The item definition needs to be considered first to determine whether the item is applicable. If not, a score of 1 (absent) should be assigned. Any evidence suggesting the item is present should prompt a score of 2 (minimal) or higher. Particularly when determining the highest score that applies (see below), efforts should be made to not reinterpret the wording, and "impressionistic" scoring should be avoided. Terms involving "and" or "and/or" should be closely attended so as to ensure that all necessary elements are present before assigning or eliminating a score from consideration.

Second principle—Always give the highest rating that applies. Very often, raters are faced with ambiguity. It might be that the answers to queries are unclear or that the information available suggests that more than one score may be applicable. A simple solution—and a "convention" frequently applied for other instruments—is to "rate up" when more than one score might be applicable. For the PANSS, a somewhat different approach is mandated, and instead of arbitrarily moving to select a score, raters should instead always give the highest score that applies based on the available information. For example, if a patient clearly meets the criteria for a score of 3 (mild) and also for 4 (moderate) on any item, as long as all the necessary criteria for both items are met, then the patient should receive a score of 4 (moderate). In the same vein, if a patient almost meets the criteria for a score of 4 (moderate), but is clearly missing some key component, then a score of 4 (moderate) cannot be assigned.

Third principle—Always consider the reference period and time frame. Some patients are not always clear about the time frame under examination during an assessment. Typically, the PANSS is rated based on a "past week" reference period (i.e. the ratings are based on the most severe phenomenon for a given item in the past week). It is worth noting, however, that certain items based solely on nonverbal symptoms during the interview, such as Item N1 (blunted effect), will be rated based on the presentation the rater can observe during the interview. Patients might describe a wide range of experiences during the course of an assessment—including some that occurred more than one week ago. While that might reveal beliefs or ideation that is, effectively, still present, many time-delimited phenomena might not be impacted. For example, Item P7 (hostility) would not be directly impacted by a fight that the patient had four weeks ago when using the standard past week reference period.

Fourth principle—Use all available information for rating, as long as it meets the basis for rating. Instruments developed for other disorders sometimes assume a linear progression with discrete sections compartmentalized by scale item. While the Structured Clinical Interview-PANSS (SCI-PANSS) does have some relatively discrete components, it is more likely that information relevant to rating different items may be presented at any time, possibly even well after the section on an item has been completed. Patients might also give conflicting information at different points during an interview, denying a symptom initially and then endorsing it later. While it is difficult to anticipate every combination of presentations or endorsements, raters should avoid assigning item scores during the interview and should instead wait until the assessment is complete and all necessary information (including informant data) is collected. At the conclusion of the assessment, all information that is relevant and meets the basis for ratings should be taken into account in the final determination.

Notably, there are several controversies that have arisen over the years with regard to the proper use of the PANSS. While the following items do not comprise an exhaustive list, they still highlight some of the challenges that raters should consider and develop techniques and strategies to address.

Is collateral (informant) information required to rate the PANSS? Two items in the PANSS (N4 and G16) are rated solely on the basis of information meant to be gathered from an informant such as a caregiver or a treating clinician who has had significant contact with the patient during the reference period. It is sometimes challenging to obtain sufficient information to cover all of the required areas, but raters are first instructed to do their best to obtain the necessary information from a third party. In the absence of any available independent person to query, the rater may use records of various sorts in order to gain insight into behaviors during the past week.

Is adherence to the SCI-PANSS necessary or is a general clinical psychiatric interview sufficient to obtain information for the purpose of rating? Most clinical trials now mandate the use of the SCI-PANSS. Lindstrom[3] and others[4] have demonstrated that high reliability can be generated between raters using the SCI-PANSS.[1] While the SCI-PANSS could be improved upon—and could be in future iterations—it is necessary to have a standardized approach to assessment across visits, patients, and investigators so as to help improve reliability. Additionally, the SCI-PANSS is designed to help ensure that all necessary domains of inquiry are addressed. It is important, however, to remember that the SCI-PANSS is intended to be used as semi-structured interview guidelines rather than a rigidly conducted script. Rewording, rephrasing, and other techniques to help improve patient comprehension can and should be engaged when applicable. Additionally, there might be instances in which it is beneficial to change the order of the questions. For example, a disorganized and challenging patient might spontaneously begin talking about hallucinatory experiences. A rater might then determine that it is clinically advisable to take advantage of the opportunity to explore this symptom further rather than attempting to redirect the interview at that point.

Is it necessary to use the anchoring points if the patient is quite severe across an entire domain (e.g. positive symptoms)? Less experienced clinicians and raters are often over-impressed by psychotic symptoms and appear to rely less on the anchor points in these instances. While it is tempting to "save time" by assigning blanket scores for items impressionistically, such an approach fails to meet the standards for reliable use of the PANSS. Raters are urged to carefully reach each item and assign the highest score that applies on the basis of the written anchors.

In cases in which the local definition of an item/concept differs from the one shown in the PANSS rating criteria, may the local alternative be substituted? Different disciplines and fields of study can variably define common concepts (e.g. delusions). In clinical practice, these approaches might have significant value to treatment of patients in a local context; for example, if a culturally influenced explanation of a symptom that is acceptable to the patient and his/her family needs to be explored and acknowledged by the treating clinician to facilitate communication and adherence with treatment, then this is of great value to all stakeholders in that context.[5] However, within the confines of a clinical trial, particularly one that is multi-site and/or global in nature, the need for standardization across visits, sites, and regions for the purposes of research necessitates that all raters adhere to the common definitions of terms without substitution.

IMPLEMENTATION OF TRAINING

Traditionally, rater training for the PANSS involved raters attending an investigator meeting for each clinical trial, where they would sit classroom style, listen to a slide-based lecture, view videotaped interviews, and rate them through an audience response system. Outlying scores were discussed with the goal of optimizing inter-rater reliability. Certification was based on scoring an agreed-upon percentage of items with fidelity to the "gold standard." At a mid-study investigator meeting intended to prevent rater drift, raters would review a slide lecture and rate an additional videotaped interview, and were remediated if their scores were outside the "gold standard."[6]

Limitations of traditional training. Such methodologies were capable of achieving and maintaining high levels of reliability and have effectively remained unchanged since the original Phase III studies of risperidone in the 1990s.[7] However, the limitations of this methodology have become apparent and are as follows: 1) raters working on multiple trials are sometimes subjected to repetitive training that does not take their individual issues in PANSS rating into account; 2) rating a videotaped interview does not address the correct assessment technique and the ability to elicit information from a psychotic patient; 3) training should be relevant and individualized to the specific clinical trial; and 4) a rater's behavior in the laboratory of an investigator meeting does not necessarily reflect the rater's behavior while at his or her site rating patients.[8]

Interactive training. PANSS training is rapidly evolving to address the above issues. Increasingly, traditional, passive, classroom-style training is being replaced with interactive, case-oriented methods that require active participation from investigators. For example, in the "roundtable approach," investigators are organized in small groups, often by site and nationality. Instead of a long repetitive lecture, there is a short review of the basic principles of rating followed by case discussions. Within each group, raters come to a consensus with their colleagues from their sites and countries. This synchronizes a rating methodology within a site and prevents "noise in the ratings" when raters cross-cover for each other. The session is moderated by an appropriately qualified trainer who is capable of synthesizing the various points of view and who is tasked with ensuring compliance to core principles and gold-standard approaches. There are many variations in this methodology but they share the concepts of active participation and consensus-building to replace passive listening.

In the past, the centerpiece of training for both beginner and advanced raters were lengthy, item-by-item ratings of full, unselected PANSS interviews. The current trend for experienced raters is to teach with shorter vignettes targeting relevant areas of study design, such as the population under study (e.g. acutely psychotic, prominent positive symptoms, predominant negative symptoms, stable, treatment resistant), change from baseline, and difficult to rate symptoms.

Assessment technique. Interview skill assessment and feedback has become integral to PANSS training and addresses the ability of the rater to probe the population under study sufficiently so as to distinguish among the anchor points of each item in a neutral manner unlikely to induce a placebo response. This is most effective when using highly trained live actors who challenge the investigator with scripted foils.

Certification procedures. In the past, certification to administer the PANSS was commonly based on the successful rating of a videotaped PANSS interview. However, this is a passive procedure that fails to assess the investigator's ability to deliver a thorough and unbiased interview. It is critical to standardize both the interview technique and measurement skills. A newer procedure for certification is to require candidates to successfully interview and measure the symptom severity of highly synchronized actors portraying patients with psychotic disorders. The use of quantified approaches to the evaluation of interview technique has been linked with data quality and signal separation, making this "active" evaluation a more relevant and meaningful approach to certification.[1]

Videotaped interviews are more commonly used than actors to evaluate assessment technique and scoring, in part because video recording is more resource-intensive than training and synchronizing actors in multiple languages and bringing them to investigator meetings. For the most part, raters with sufficient credentials and experience administering the PANSS to the population under study are certified if they meet certain standards of accuracy and precision with their measurement of the individual PANSS items and the PANSS total, based on both gold standards and statistical outliers. To accelerate the rater approval process, decisions regarding success or failure of the candidate as well as remediation may be delivered at the investigator meeting. Like any assay, the measurement of psychotic symptoms must be periodically recalibrated. Intra-rater and inter-rater reliability should be assessed and remediated regularly.

IMPACT OF TECHNOLOGY

Technology has provided vibrant, efficient alternatives to expensive, potentially inefficient in-person, multi-country investigator meetings. Initial training, as well as mid-study refresher training, may occur by use of "live" web conferencing, essentially recapitulating the interaction of an investigator meeting, or in an "on-demand" manner, either online or application-based.[9] Adaptive and risk-based methods may be applied to individualize PANSS training to triage a rater to more basic or advanced

nuanced curriculums or to steer the training toward specifc areas for improvement. Avatars can be programmed with decision tree logic to serve as subjects for interview skills training. Virtual reality may be used to create a realistic assessment environment. All these technologies, and more to come, might transform traditional training and make it more useful, practical, and effective in years to come.

Use of electronic clinical outcome assessment (eCOA). Another means by which newer technologies can bolster PANSS training and data quality is use of eCOA. Platforms utilizing this methodology can assess ratings for logical inconsistencies among PANSS items and between the PANSS and other scales and alert the investigator before data submission. The investigator has the option to reevaluate their rating or to maintain the original scores. eCOA also permits additional alerts and reminders to be made to the rater. For example, the PANSS rater may be prompted to include informant information when appropriate or to periodically remind the subject of the reference period. Notes to support the choice of anchor point might be required. This technology was positively received by both patients and caregivers, with minimum modification requests.[10]

The capacity for audio/video recording of SCI-PANSS interviews can be embedded in the eCOA platform to facilitate deeper independent review of visits, either through an *a priori* plan (e.g. evaluation of every rater's first assessment) or via a risk-based approach using inconsistencies detected within PANSS data to "flag" an evaluation for review. Early detection and remediation of these data flaws is critical for study success and to prevent "rater drift."[11] Continual evaluation of the quality of a site's interviews and ratings and retraining as necessary should continue throughout all phases of the trial, just as any assay would be repeatedly monitored and recalibrated.

EVALUATION OF NEWER TRAINING AND DATA MONITORING PROCEDURES

There have been a number of solutions to managing rater drift during clinical trials. Remote, independent rating,[12] smaller trials with more experienced rater cohorts,[13] and a number of in-study techniques that utilize the internal logic of instruments like the PANSS have gained attention in the last decade.[3,14,15] The latter technique uses algorithms to generate flags for what is often referred to as a risk-based approach to monitoring in-study data. Algorithms can consist of logical binary or factorial relationships between one or more scale items or more sophisticated statistical techniques that leverage large clinical trial datasets with known outcome parameters. For the purposes of this article, we will limit our discussion to the sorts of binary and factorial relationships that exist within the PANSS and how these can be used to generate flags. For example, if a rater scores at the level of 7 on Item P5 (grandiosity) and then scores Item P1 (delusions) at the level of 1, this would generate a flag. This is because at the level of 7 on P5, we expect significant and pervasive grandiose delusions and, if that is the case, then the P1 should receive a similarly severe rating. While this is an extreme example (and usually related to the rater's reluctance to "double rate" the same symptom) it serves to illustrate the essential idea that the instrument relationships themselves can show us where there is a high risk for error. Another illustration comes in the form of the Marder[16] five-factor model for the PANSS (though some dispute this factor solution[8]); in such frameworks, the expected correlations between items that represent factors can be used to detect aberrant presentations and potential risk.[5,11] For example, if we think about the negative factor that includes N1 to N4, N6, and G7 and we expect that these will be predictably correlated (within certain severity ranges), we can identify risk when one or more correlation fails to agree with the identified matrix.

How are these risks are dealt with? Is it actually rater error that is present? Or is it simply a somewhat unusual patient presentation? Intervention methods differ and depend on who is leading the data-monitoring effort, but if actual rater error is responsible, this is the point at which a targeted training event takes place. It must be emphasized here that an expert clinician with a very clear understanding of the scale and the patient population must complete the training. This in-study targeted training is essential in arresting rater drift and reducing the impact of non-informative data (i.e., data that contribute little to the goal of the study but increase variance and thus the ability to detect the signal where it exists). This method has proved cost-effective, and the targeted nature of intervention requires fewer resources than interval retraining (e.g., training done every 3–6 months) for the full cohort of raters. More importantly, the reduction in non-informative data can make the difference between a failed or negative trial and one that is positive.

Prospective, adequately controlled comparisons of methodologies for rater training or in-study data quality monitoring coupled with remediation are rare because sponsors are reluctant to vary methodologies within a clinical trial. The comparison of methodologies across trials is complicated by multiple uncontrolled differences in trial characteristics. That said, used in parallel, the methodologies are complimentary and can reinforce the four principles critical to obtaining reliable and valid data for the duration of a trial. Although the results must be evaluated carefully, comparisons of inter-rater reliability, nonspecific variance, placebo response, and drug-placebo differences across trials using different methodologies can be informative, if not definitive.[15] Newer interview training and scale rule training techniques can be evaluated against in-study metrics based on error rates detected by data analytics as well as via an external expert review of recorded patient interviews. The independent review of patient interviews is highly recommended for all clinical trials. It has been demonstrated that interviews that are recorded and reviewed have PANSS scores that align better with the scale requirements.[17]

CONCLUSION

PANSS rater training has become a standard component of most clinical trials, but true standardization with respect to the exact approaches, techniques, and standards remains elusive. For clinical trials using the PANSS, it is strongly advised that the training program incorporates the core principles described in this article and advocated by the author of the PANSS. Where possible, we also further recommend the following:1) Favor active learning techniques over passive ones, particularly for experienced clinicians and raters with meaningful prior experience using the PANSS. While some raters have persistent idiosyncrasies in their approaches

to the use of the scale, active approaches will be far more effective in highlighting these issues and enabling retention of new concepts and information; 2) evaluations of inter-rater reliability should include a videotaped interview or evaluation of a standardized subject/volunteer; most optimally, certification will involve an assessment of interview technique as well as inter-rater reliability to ensure that all prospective raters are capable of conducting an evaluation that strikes the proper balance of adherence to the interview guide and maintenance of flexibility and clinical research rapport; and lastly 3) following initial training, quality assurance approaches should include ongoing evaluation of data and assessment technique, are employed. Where possible, technology can and should be used to help facilitate these processes. Whether utilizing eCOA to replace paper with electronic forms or driving "targeted calibration" through an analysis of data in-study, a dynamic approach to ensuring inter-rater reliability will help to guarantee that core principles are applied rigorously throughout the study.

REFERENCES

1. Sajatovic M, Gaur R, Tatsuoka C, et al. Rater training for a multi-site, international clinical trial: What mood symptoms may be most difficult to rate? *Psychopharmacol Bull.* 2011;44(3):5–14.
2. Kay SR, Fiszbein A, Opler LA. The Positive and Negative Syndrome Scale (PANSS) for schizophrenia. *Schizophr Bull.* 1987;13(2): 261–276.
3. Lindström E, Wieselgren IM, von Knorring L. Interrater reliability of the Structured Clinical Interview for the Positive and Negative Syndrome Scale for schizophrenia. *Acta Psychiatr Scand.* 1994;89(3):192–195.
4. Knorring LV, Lindström E. Principal components and further possibilities with the PANSS. *Acta Psychiatrica Scandinavica.* 1995;92(s388):5–10.
5. Napo F, Heinz A, Auckenthaler A. Explanatory models and concepts of West African Malian patients with psychotic symptoms. *European Psychiatry.* 2012;27 Suppl 2:S44–S49.
6. Müller MJ, Wetzel H. Improvement of inter-rater reliability of PANSS items and subscales by a standardized rater training. *Acta Psychiatr Scand.* 1998;98(2):135–139.
7. Leucht S, Kane JM, Kissling W, et al. What does the PANSS mean? *Schizophr Res.* 2005;79(2-3):231–238.
8. van der Gaag M, Cuijpers A, Hoffman T, et al. The five-factor model of the Positive and Negative Syndrome Scale I: confirmatory factor analysis fails to confirm 25 published five-factor solutions. *Schizophr Res.* 2006;85(1-3):273–297.
9. Kobak KA, Feiger AD, Lipsitz JD. Interview quality and signal detection in clinical trials. *Am J Psychiatry.* 2005;162(3):628.
10. Tolley C, Rofail D, Gater A, Lalonde JK. The feasibility of using electronic clinical outcome assessments in people with schizophrenia and their informal caregivers. *Patient Relat Outcome Meas.* 2015;6:91–101.
11. Kobak K, Opler M, Engelhardt N. PANSS rater training using Internet and videoconference: results from a pilot study. *Schizophr Res.* 2007;92(1-3):63–67.
12. Shen J, Kobak KA, Zhao Y, et al. Use of remote centralized raters via live 2-way video in a multicenter clinical trial for schizophrenia. *J Clin Psychopharmacol.* 2008;28(6):691–693.
13. Alphs L, Benedetti F, Fleischhacker W, Kane J. Placebo-related effects in clinical trials in schizophrenia: what is driving this phenomenon and what can be done to minimize it?. *Int J Neuropsychopharmacol.* 2012;15(7):1003–1014.
14. Rabinowitz J, Schooler N, Anderson A, et al. Consistency checks to improve measurement with the Positive and Negative Syndrome Scale (PANSS). *Schizophr Res.* 2017 Mar 8. pii: S0920–9964(17)30141-X.
15. Daniel D, Kalali A, West M, et al. Data quality monitoring in clinical trials: Has it been worth it? an evaluation and prediction of the future by all stakeholders. *Innov Clin Neurosci.* 2016;13(1–2): 27–33.
16. Marder SR, Davis JM, Chouinard G. The effects of risperidone on the five dimensions of schizophrenia derived by factor analysis: combined results of the North American trials. *J Clin Psychiatry.* 1997;58(12):538–546.
17. Kott A, Daniel DG. Effects of PANSS audio/video recordings on the presence of identical scorings across visits. *Eur Neuropsychopharmacol.* 2015;25:S543–S544. ICNS