# Evaluating the Impact of Rater Change on PANSS Score Change: An Item Level Analysis

Verma, P[1], Micaletto, M[1], Laird, R[1], Hong, B[1], Feaster, T[1],Tackett, ZC[1], Machizawa, S[1], Zhou, W[1], Crittenden-Ward, K[1], Daniel, DG[1], Kott, A[1]

[1] Signant Health, Reading, UK

## ABSTRACT

The PANSS is a commonly employed scale to assess symptoms of schizophrenia requiring clinical evaluation of data collected through clinical interviews with participants and caregivers. This poster explores the impact of PANSS rater change on PANSS item-level variability by comparing score changes between ratings made by the same rater vs. ratings made by different raters. Overall median score change was significantly different between the two groups. The majority of PANSS items showed susceptibility to rater change as evidenced by a significantly higher variability in scores. Only eleven PANSS items resulted as having non-significant score changes even in instances where the raters were different at subsequent visits. This research carries important implications for sponsors considering using the PANSS for inclusionary and/or endpoint assessment in clinical trials. It also highlights the need for robust rater training and endpoint reliability programs in clinical trials.

## BACKGROUND

The Positive and Negative Syndrome Scale (PANSS) is a commonly employed scale to assess symptoms of schizophrenia. It requires clinical evaluation based on data collected through clinical interviews (with participant and caregiver) and historical knowledge of both the participant and his/her condition. This makes the PANSS a highly subjective scale as raters must collect, consider and calibrate scores based on their individual clinical judgement and knowledge of the population at hand, and then assign a rating that best reflects their overall clinical impression of the participant whilst adhering to scale guidelines. Therefore, idiosyncrasies in rating, interview style and scale interpretation can all lead to poor intra- and interrater reliability. Robust training combined with calibration of raters' interview skills and rating techniques is essential to collecting valid and reliable data.

Previous analyses have shown rater change is associated with increased within-subject variability (Kott, Wang & Daniel, 2019). The results showed that the increase was not consistently seen in all 5 PANSS factors and 2 PANSS factors - PANSS Uncontrolled / Excitement factor and Positive factors did not show increased within-subject variability. The authors hypothesized, that increased variance observed in PANSS Negative, Disorganized and Anxiety and Depression factors may be linked to the fact that positive symptoms are more easily observed and scored. However, this previous research conducted analysis at a factor level, by grouping several PANSS items together.

Our current research focuses on how rater change impacts PANSS item-level score changes.

## METHODS

PANSS data were evaluated from 4 multi-national, double-blind, placebo controlled clinical trials in schizophrenia.

In each of the 4 multi-national trials, raters were trained and certified on the PANSS. All PANSS raters met experience criteria (educational degree and scale) for the respective protocol and were qualified on a PANSS scoring exercise and interview skills assessment prior to administering the scale in a study. There were 350 distinct raters who administered the PANSS at a Screening visit with 10,100 post-screening visits. Of the post-screening visits, 93.0% (9,397) had the same rater from Screening and 7.0% (703) had a different rater.

In this analysis, PANSS score change for each item was compared between groups (with rater change and without rater change).

## RESULTS

A Shapiro-Wilk test for normality on the PANSS score change found the distribution to be significantly different from a normal distribution (all ps < 0.0001). A Mann-Whitney test was conducted for each item to determine the influence of rater change on score change.

The Mann-Whitney test results were broken down into tables with the most significant items in Table 1 and the least significant items in Table 4. Table 1 shows most significant differences in score changes were found with the following items: G10 Disorientation, N7 Stereotyped Thinking, and G12 Lack of Judgment and Insight ($p < 0.0001$). Table 2 and Table 3 displays the items that were significant but not as significant as those in Table 1. Table 4 shows the items that had no significant difference of rater change on score change ($p > 0.05$). More than half of the PANSS item showed the influence of rater change to be significant.

Figure 1 shows the score change distributions between the rater change and the no rater change groups for the three key PANSS items noted in Table 1 (G10, N7 and G12). The score change distributions appear to be similar, lending more credibility to the Mann-Whitney results. The Mann-Whitney tests the hypothesis of whether the two groups have the same continuous distribution by checking for similar shape and similar location. By confirming the two groups have the same shape (or the same distribution for a range of score changes), we can be confident the significance of the Mann-Whitney results is due to location (or mean score change difference) between the two groups.

Figure 2 shows a graph of the score change difference between the rater change and the no rater change groups at an item level. Across all PANSS items, with the exception of items G14 Poor Impulse Control, G1 Somatic Concern, and G2 Anxiety, the rater change group displayed a greater score change from the previous visit than the no rater change group on average.

Statistical significance was defined as $p<0.05$. All analyses were performed using R v 3.6.1.

**Table 1 - Significant Score Differences by Item between the Rater Change vs. No Rater Change Group (PANSS Items with p value <0.0001)**

| item | change (Rater change - No rater change) | p_adj | sig |
|---|---|---|---|
| G12 | 0.084 | 1.03E-05 | *** |
| G10 | 0.099 | 1.17E-06 | *** |
| N7 | 0.086 | 6.92E-06 | *** |

**Table 2 - Significant Score Differences by Item between the Rater Change vs. No Rater Change Group (PANSS Items with 0.0001< p value <0.01)**

| item | change (Rater change - No rater change) | p_adj | sig |
|---|---|---|---|
| G5 | 0.053 | 4.22E-03 | ** |
| G13 | 0.074 | 1.19E-03 | ** |
| P2 | 0.067 | 9.82E-04 | ** |
| G9 | 0.058 | 1.49E-03 | ** |
| N1 | 0.079 | 1.87E-04 | ** |
| G15 | 0.075 | 1.49E-04 | ** |
| P5 | 0.074 | 1.98E-03 | ** |
| N3 | 0.084 | 8.74E-04 | ** |
| G3 | 0.076 | 2.67E-04 | ** |
| P7 | 0.045 | 9.86E-03 | ** |
| P4 | 0.062 | 5.32E-04 | ** |
| G6 | 0.08 | 1.11E-03 | ** |

**Table 3 - Significant Score Differences by Item between the Rater Change vs. No Rater Change Group (PANSS Items with 0.01< p value <0.05)**

| item | change (Rater change - No rater change) | p_adj | sig |
|---|---|---|---|
| G8 | 0.036 | 3.25E-02 | * |
| N4 | 0.061 | 1.08E-02 | * |
| N5 | 0.06 | 1.31E-02 | * |
| N6 | 0.057 | 1.01E-02 | * |
| P6 | 0.045 | 2.78E-02 | * |

**Table 4 - Significant Score Differences by Item between the Rater Change vs. No Rater Change Group (PANSS Items with p value >.05)**

| item | change (Rater change - No rater change) | p_adj | sig |
|---|---|---|---|
| N2 | 0.027 | 3.28E-01 | ns |
| P1 | 0.03 | 2.10E-01 | ns |
| G14 | -0.029 | 3.54E-01 | ns |
| G11 | 0.026 | 2.44E-01 | ns |
| G16 | 0.025 | 3.46E-01 | ns |
| G7 | 0.03 | 1.55E-01 | ns |
| P3 | 0.009 | 4.12E-01 | ns |
| G4 | 0.008 | 4.67E-01 | ns |
| G1 | -0.019 | 9.57E-01 | ns |
| G2 | -0.011 | 7.04E-01 | ns |

**Figure 1 - Percentage of the Score Change Distribution for the Top 3 Significant PANSS Items occurrence**

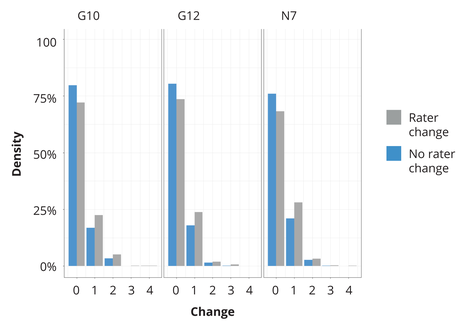
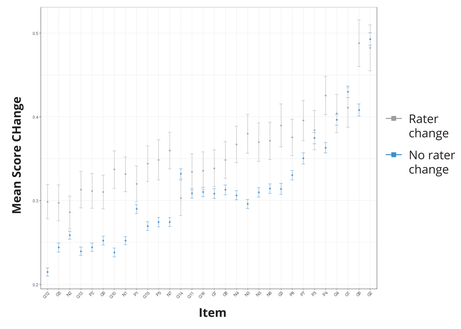



**Figure 2 - Mean Score Change by Item between the Rater Change vs. no Rater Change Group**




## DISCUSSION

Overall,19 out of 30 of the PANSS Items were significantly affected by rater change from previous visit. The three most significant items impacted were those that relied on observation during the interview, clearly highlighting how idiosyncratic evaluation can impact PANSS ratings.

There appear to be a number of PANSS items which are more resilient to rater change (Table 4). Seven of these eleven items are based, to an extent, on reports from caregivers. This suggests that items evaluated with consideration for "external" reports may be less subject to variation due to rater change.

The results indicate rater change can introduce significant noise into PANSS assessments and this could have a potentially significant impact in clinical trials using this scale as an inclusionary and/or endpoint measure. Whilst PANSS items relying on external report appear more resilient to rater variability, this research highlights the paramount need to ensure PANSS raters are well-trained and well-calibrated within any clinical trial and rater change is kept to a minimum or eliminated when possible.

Potential recommendations for new trials include: robust rater training and ongoing calibration through life cycle of the clinical trial and endpoint data monitoring programs to provide surveillance and safeguard data quality.

Further research should explore additional confounding variables that may affect PANSS ratings such as time of day of assessment or duration of assessment.

## REFERENCES

Kott, Alan; Wang Xingmei; Daniel David (2019): Exploring the association of PANSS rater change with extreme within-subject variability in schizophrenia clinical trials. Poster presentation. International Society for CNS Clinical Trials and Methodology (ISCTM). Washington, USA, 19-21/2/2019.

*Financially supported by Signant Health.*

*Presented at the ISCTM 2020 Autumn Virtual Conference, September 21 - 25, 2020.*

SIGNANT HEALTH