Check for updates

STUDY PROTOCOL

Applying advanced psychometrics to antipsychotic treatment trials: A protocol comparing effect estimates

[version 1; peer review: 1 approved, 2 approved with reservations]

Timo Schurr[1], Fiona Boland[2], David Byrne[1], Frank Doyle[1]

[1]Department of Health Psychology, School of Population Health, RCSI University of Medicine and Health Sciences, Dublin, Ireland
[2]Data Science Centre, School of Population Health, RCSI University of Medicine and Health Sciences, Dublin, Ireland

v1 **First published:** 17 Jun 2026, **6**:211
https://doi.org/10.12688/openreseurope.23623.1
Latest published: 17 Jun 2026, **6**:211
https://doi.org/10.12688/openreseurope.23623.1

Abstract

Background

Antipsychotic trials in schizophrenia quantify treatment efficacy using clinician-rated scales such as the Positive and Negative Syndrome Scale (PANSS) and Brief Psychiatric Rating Scale (BPRS). If these scales include redundant or weakly informative items, measurement noise can blur drug-comparator differences that may dilute real effect sizes. Modern psychometric methods can refine and re-weight symptom scores that may improve precision and potentially increase detectable treatment effects, although it is unclear which approaches might be best.

Aims

To 1) determine best psychometric models for the PANSS and BPRS, and 2) test whether psychometrically-informed scoring of PANSS/BPRS alters estimated antipsychotic effect sizes compared with conventional total-score analyses, and 3) to explore whether any changes differ across symptom domains and by sex.

Method

We will conduct secondary analyses of anonymised individual participant data from randomized Phase II–IV schizophrenia trials accessed via secure data-sharing platforms (YODA Project and Vivli). Item-level PANSS/BPRS data will be harmonised at baseline and outcome. Psychometric models will be derived using confirmatory factor analysis, item response theory, and network analysis. Stability and invariance analyses will be conducted, examining covariates such

Open Peer Review

Approval Status ? ✓ ?

	1	2	3
version 1 17 Jun 2026	? view	✓ view	? view

1. **Robin Emsley**, Stellenbosch University, Stellenbosch, South Africa
2. **Yaoyao Sun**, Peking University Sixth Hospital, Peking University Institute of Mental Health, Beijing, China
3. **Felice Iasevoli**, University School of Medicine of Naples Federico II, Naples, Italy

Any reports and responses or comments on the article can be found at the end of the article.

as age and sex. Antipsychotic efficacy will be estimated using mixed-effects participant-level models, expressed as Cohen's d standardised mean differences (SMDs) for both conventional and psychometrically-informed outcomes. The primary outcome will be the difference in SMDs between conventional and psychometrically-informed effect sizes, summarised across trials.

Conclusions

This study will determine whether best practice psychometric refinement of existing measures meaningfully changes antipsychotic effect size estimates, informing outcome measurement choices and interpretation of trial results.

Keywords
schizophrenia, antipsychotics, panss, bprs, effect size, psychometrics, confirmatory factor analysis, item response theory, network analysis, individual participant data

This article is included in the Horizon Europe gateway.

This article is included in the Marie-Sklodowska-Curie Actions (MSCA) gateway.

Corresponding author: Timo Schurr (timoschurr@rcsi.com)

Author roles: Schurr T: Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Boland F**: Conceptualization, Methodology, Project Administration, Software, Supervision, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Byrne D**: Conceptualization, Methodology, Project Administration, Software, Supervision, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Doyle F**: Conceptualization, Funding Acquisition, Methodology, Project Administration, Software, Supervision, Validation, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101209817 (rEVealIng Drug Effects when applyiNg psychomeTrics In schizophreniA triaLs [EVIDENTIAL]).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2026 Schurr T *et al*. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Schurr T, Boland F, Byrne D and Doyle F. **Applying advanced psychometrics to antipsychotic treatment trials: A protocol comparing effect estimates [version 1; peer review: 1 approved, 2 approved with reservations]** Open Research Europe 2026, **6**:211 https://doi.org/10.12688/openreseurope.23623.1

First published: 17 Jun 2026, **6**:211 https://doi.org/10.12688/openreseurope.23623.1

Background

Psychometric assessment and its relevance for estimating treatment effects in schizophrenia research

Multi-item psychometric scales are fundamental to clinical and psychiatric research and central to the assessment of symptom severity, treatment response and prognosis in schizophrenia. The Positive and Negative Syndrome Scale (PANSS[1]) and the Brief Psychiatric Rating Scale (BPRS[2]) are the most frequently used clinician-rated instruments to assess symptoms. Underpinning treatment guidelines and regulatory approval,[3] both scales are also widely used as primary outcome measures in randomised clinical trials (RCTs).

Over several decades, there has been substantial effort to evaluate the psychometric properties of both these instruments. In addition to approaches spanning classical test theory (e.g. confirmatory and exploratory factor analysis), more sophisticated statistical methods like Rasch-modelling and network analysis have been used in order to attempt to clarify the symptom structure, improve measurement precision or identify poorly performing items.[4–8] This work has produced a large collection of evidence with differing or partially contradictory conclusions with respect to item functioning or scale structure and dimensionality. Most psychometric theories emphasise that validity is not an inherent property of an instrument, but rather that it depends on how well it measures the intended construct in a given population or context.[9,10] Hence, when applied across different populations or interventions, continuous evaluation of psychometric scales using.[11,12] However, it is not clear to what extent these continued psychometric refinements meaningfully improve conclusions drawn from studies using these instruments.

In principle, psychometrical optimization should increase statistical power by reducing measurement error, therefore, yielding more precise effect size.[13,14] However, there is no consensus in the field that these expectations are empirically met, since a growing body of methodological literature shows that psychometrically refined scores are often.[15,16] The claim of 'improved measurement' following psychometrical refinement therefore.[11] Current research in the field of depression has begun to fill this gap by linking psychometric analyses to the effect size estimation in randomised controlled trials (RCTs). Through combining individual participant data (IPD) from several RCTs, these investigations were able to compare treatment effects derived from psychometrically-informed models (from factor analytic, item response theory or network analysis models), to those with conventional total scoring. Results indicated that factor-analytical approaches could be increase estimated antidepressant treatment effects not universally improve outcome measurement the importance of careful interpretation of clinical relevance. Crucially, this emerging evidence reframes psychometrical optimization as a means to evaluate whether measurement choices substantially influence the estimation of treatment effects. From this perspective, it is no longer the question of how psychometric models best depict symptom structures, but whether conclusions drawn from such models meaningfully inform science, clinical practice or policy.

Despite continuous efforts in pharmacological development, reported effect sizes remain modest in schizophrenia trials.[16,17] Although likely, it is still debatable whether shortcomings in the outcome measurements contribute to this pattern.[6] Despite extensive analysis of the psychometric properties of the PANSS and BPRS, no large-scale efforts have been made to assess whether psychometric refinements alter the estimated treatment effects in schizophrenia trials, nor whether such effects differ between demographic subgroups (e.g. sex) or symptom domains. This suggests the need for a systematic, large-scale assessment of psychometric approaches in schizophrenia research, explicitly examining their implications for effect size and outcome measurement.

The present protocol therefore describes how the EVIDENTIAL project[18] will address this gap in knowledge. This project will adopt similar methods to those described by Doyle and colleagues,[19] replicating and extending these to schizophrenia trials. The primary aim of the EVIDENTIAL project is to comprehensively examine the properties of the BPRS and PANSS to determine whether psychometric refinement might meaningfully alter schizophrenia symptom structure or performance, and more crucially, whether these changes might influence estimates of treatment efficacy examined in subsequent analyses. To achieve this, we have three objectives, corresponding to three research stages:

- To determine the best psychometric models for the PANSS and BPRS
- To test whether psychometrically-informed scoring of PANSS/BPRS alters estimated antipsychotic effect sizes compared with conventional total score analyses
- To explore whether any changes differ across symptom domains and by sex.

Methods

Study design

This study is a secondary IPD re-analysis of randomized controlled trials of antipsychotics in schizophrenia, schizoaffective or schizophreniform disorders. Item-level symptom ratings from the PANSS and BPRS will be collected and psychometrically analysed at baseline and outcome assessment. Reporting will follow STROBE[20] principles, and any deviations from the prespecified plan will be documented.

Setting, participants and eligibility criteria

This study is open to randomized controlled multi-arm trials (Phase II-IV) comparing the effects of pharmacological treatments with antipsychotics to placebo (primary outcome) or any comparator (sensitivity analysis). Individual trial size is not an eligibility criterion. Eligible trials will include adult participants diagnosed with schizophrenia, schizoaffective or schizophreniform disorders. Trials must include PANSS and/or BPRS individual participant data at baseline and outcome. Studies meeting the criteria described in Table 1 have been requested from vivli.org (Vivli) and yoda.yale.edu (YODA Project). Requested studies from the YODA Project are solely sponsored by Johnson & Johnson. In addition, Vivli provides data from the following sponsors: Abbott, AbbVie, Janssen, Johnson & Johnson, Lilly, Otsuka. Previous data checks suggests that we should expect to obtain a pooled sample of ~31 000 participants from 77 studies provided on the Vivli platform (Table 2).

Outcome measures: PANSS and BPRS

The PANSS[1] is one of the most widely used clinician-rated outcome measures in schizophrenia research. It comprises 30 items, each rated on a 7-point anchored severity scale (1 = absent to 7 = extreme), which is commonly summarised as three symptom groupings: Positive (7 items), Negative (7 items), and General Psychopathology (16 items), together forming a total score ranging from 30 to 210. Despite its extensive use, heterogeneity in reported factor solutions and item functioning has prompted concerns about the PANSS' psychometric robustness and its sensitivity to treatment-related change under some conditions.[6,23]

The BPRS is another long-established clinician-rated scale frequently used in psychopharmacology research to quantify a broad range of psychiatric symptoms (e.g., thought disturbance, affective symptoms, activation, hostility).[2,24] We will utilize the BPRS version comprising 18 items, each rated on a 7-point severity scale (1 = not present to 7 = extremely

Table 1. Inclusion/exclusion criteria for selecting trials and participant data.

Variable	Inclusion criteria
Study Type	Randomized controlled multi-arm trial
Data Type	Individual Participant Data
Phase	II, IIa, IIb, III, IIIa, IIIb, and IV
Population/Diagnostic Criteria	Participants with Schizophrenia Spectrum Disorders, aged 18 years and older
Treatment	Antipsychotic medication
Comparator	*Placebo* (to establish efficacy) *Active Comparator* (to establish baseline psychometric model)
Active Comparator	*Primary* (standard)*:* first-generation vs second-generation antipsychotics (optionally adding D2 partial agonists as a third category where feasible). *Exploratory:* receptor-based taxonomy as proposed by McCutcheon et al.[21]
Outcome	Positive and Negative Syndrome Scale (PANSS) and/or Brief Psychiatric Rating Scale (BPRS)
Covariates	Sex (if available: Age)
Moderator Variable	Sex Stage of illness (sensitivity analysis)
Timing of Outcome Assessment	6 weeks, as per Younis et al.[22]*/4 weeks for sensitivity analysis*
	Exclusion Criteria
	All studies outside of the above parameters

Table 2. Expected Sample.

Database One:		**https://yoda.yale.edu/ (YODA Project)**
Number of studies requested	-	62
Number of studies by sponsor		
Janssen	-	26
Johnson & Johnson	-	36
Database Two:		**www.vivli.org (Vivli)**
Number of studies requested	-	77 (62 are overlapping with YODA)
Number of studies by sponsor		
Abbott	-	1
AbbVie	-	1
Lilly	-	9
Janssen	-	26
Johnson & Johnson	-	36
Otsuka Pharmaceuticals	-	4

severe), producing total scores that typically range from 18 to 126. As with PANSS, published work indicates that factor structures and item performance can vary across samples and settings, which is relevant when BPRS totals are used as trial outcomes.[7,8,25,26]

Statistical analysis plan

Analyses will be conducted within the secured environment of Vivli. Our approach will follow a two-stage workflow. First, we will use item-level PANSS and BPRS data to evaluate and select the best-fitting, state-of-the-art psychometric models of symptoms. Second, we will use the resulting psychometrically-informed scores to estimate antipsychotic treatment effects and compare these with effects derived from conventional PANSS/BPRS total scores, thereby quantifying whether and how psychometric modelling alters trial-level and pooled treatment effect estimates. We will follow a similar methodological approach as detailed in the publications provided by Doyle et al.,[19] regarding the exploratory and confirmatory factor analyses (EFA, CFA) and Byrne et al.,[11,13,15] regarding the item response theory framework (IRT) and network analytical approach (NA).

Stage 1: Establishing optimal psychometric models of PANSS and BPRS

Exploratory and Factor analytic approaches

Data will be randomly split into exploratory and confirmatory datasets, as recommended by Lubke and Luningham,[27] with stratification by treatment group and trial to ensure balanced representation. For stratification and subsequent subgroup/invariance checks, active antipsychotic arms will be coded using literature-standard classification (primary approach: first-generation vs second-generation antipsychotics; where feasible, a third category for D2 partial agonists may be considered). In exploratory analyses, we will additionally code active agents using McCutcheon and colleagues' data-driven receptor-based taxonomy,[21] namely muscarinic (M2-M5) receptor antagonists; dopamine (D2) partial agonists and adrenergic antagonists; serotonergic and dopaminergic antagonists; and dopaminergic antagonists. We will ascertain factor models of schizophrenia for baseline (pre-treatment), combined outcome (post-treatment; 6 weeks) and separately for placebo and treatment groups. We will further evaluate measurement invariance across major arm categories (e.g., placebo vs active treatment and, where supported, across active treatment-arm classes) within the best-fitting baseline model(s).

Parallel analysis will be applied to the exploratory datasets to determine symptom dimensionality.[28] Next, the EFA will be performed, using oblique rotation to allow for correlated symptom domains, to examine the factor structure of the exploratory dataset. Models will be refined iteratively by evaluating factor loadings below 0.3 and/or communalities below 0.4, with poorly performing items removed where justified to achieve stable and interpretable solutions.[29]

In the confirmatory dataset, CFA will be used to formally test candidate psychometric models identified during the exploratory phase. In addition to simple structures derived from EFA, established alternative models, including higher-order and bifactor representations, will be evaluated.[30] To select the optimal model of schizophrenia, the selection process

will be guided by established absolute and relative fit indices, information criteria, and internal consistency metrics, including: the Root Mean Square Error of Approximation (RMSEA; < 0.08), Standardised Root Mean Square Residual (SRMR; < 0.08), Comparative Fit Index (CFI; > 0.95), Tucker-Lewis Index (TLI; > 0.95), Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and McDonald's Omega (acceptable >0.70), following established practices.[31,32] Similar to the exploratory approach, confirmatory modelling will be performed for baseline and for the primary post-baseline outcomes. For outcome (post-baseline) measurement models, we will estimate the best-fitting model in the pooled sample and then test invariance across key arm categories (placebo vs active; and where feasible across active-arm classes as defined by the standard classification, with McCutcheon's taxonomy explored in sensitivity analyses). We will retain the best-fitting baseline and post-baseline model(s) to derive factor scores for downstream treatment-effect estimation in Stage 2.[33]

Measurement invariance, i.e. metric invariance (factor loadings are equal) and scalar invariance (intercepts are equal) across sex will be assessed, and differential item functioning examined by mean, standard error, standard deviation, factor loading scores and r^2 values. When invariance is violated, psychometrically optimized models will be examined separated by sub-group (e.g., women and men).

The outcomes of these analyses will then be used for further evaluation and discussion (descriptively) on their differences or similarities in all models and datasets. This procedure will include model structure or item functioning (i.e. dropped items, item loadings on factor). Revised sum scores based on the retained items and weighted scores will then be computed from the best performing baseline and combined outcome model for each measure. These will be used for the stage 2 analysis described in the following sections.

Item response theory approaches

Similar to the procedure described above, we will also use IRT approaches.[34,35] In IRT, each item is treated as providing information about a latent trait, and items are evaluated by how strongly they discriminate between individuals at different levels of that trait (i.e., higher latent severity increases the probability of endorsing higher response categories). Within an exploratory dataset, we will use non-parametric IRT (Mokken scaling) to assess scalability, dimensionality, and item ordering, and to identify poorly discriminating items. Following recommendations by Meijer and Baneke,[36] we will evaluate solutions across increasing lower-bound scalability thresholds (c), beginning at c = 0.01 and then stepping by 0.05 (i.e., 0.05, 0.10, 0.15 … up to 0.60), selecting the most interpretable solution. Second, using the confirmatory dataset, we will fit parametric IRT graded response models to estimate item parameters (e.g. discrimination and category thresholds) to identify the best fitting model. Hence, we will compute revised sum scores based on the retained items and generate latent trait (weighted) scores for use in downstream analyses within stage 2.[37] A detailed description of the analysis principles for the IRT approach is outlined in Byrne's publications.[11,13]

Network analytic approaches

Network analysis is a more recent approach to psychometrics, which differs from traditional theories in that it does not assume that symptoms are caused by a single underlying latent trait. Instead, NA conceptualises symptoms as mutually interacting components, represented as a graph in which nodes are symptoms and edges reflect conditional associations between symptoms.[38,39] This perspective has been proposed as a clinically intuitive complement to traditional models of psychopathology. Using best practice approaches, we will estimate empirical network structures separately for baseline (pooled across arms) and outcome data (separated by treatment arm). Network analysis will follow an exploratory graphical workflow, to identify the dimensional structure and detect redundant or weakly-performing items. As recommended by Christensen and Golino,[40] bootstrap-based stability assessment will be applied and empirical networks will be assessed in relation to the bootstrapped samples. Their work also suggests that replication metrics around the ~0.65–0.75 range mark the point at which variables begin to show reduced stability, providing a practical basis for identifying potentially unstable items.[12,40] A key goal is to identify a network structure that is stable and configurally invariant across baseline and outcome (or treatment arms). If invariance cannot be achieved, we will proceed with the network solution with fewest/least consequential instabilities. From the final network solution, we will compute (i) revised sum scores based on the retained items, and (ii) weighted "net scores" derived from network loadings/strength. A detailed description of the analysis principles for the NA approach is outlined in Byrne's publication.[15]

Stage 2: Does psychometric modelling matter? Modelling change and calculating effect sizes

In stage 2 we apply the psychometrically-derived outcomes from the stage 1 analysis to evaluate whether psychometric refinement alters estimated antipsychotic treatment effects compared to original total scores.

Primary analyses: Continuous outcomes

For each instrument (PANSS and BPRS), outcome symptom severity will be modelled as a function of treatment assignment, adjusting for baseline symptom severity and accounting for clustering of participants within trials via random intercepts, as per best practice RCT modelling.[41] Detached from the psychometric modelling approach (CFA, IRT, NA) in Model 1 the overall effects of antipsychotics versus placebo, using conventional scoring (i.e., original raw scores) will be estimated. Model 2 comprises a set of related models that mirror Model 1, where raw scores are replaced with abbreviated sum scores based on the results of the psychometric analysis from stage 1. In Model 3 psychometrically-informed (weighted) scores, that are derived from the abbreviated models will be established. Model 2 and 3 will be calculated for each psychometric method separately (CFA, IRT, NA). Models 4, 5 and 6 will extend Models 1, 2a-c and 3a-c, by additionally adjusting for participant level covariates:

- Model 1 (conventional scoring): Outcome is the raw total score with treatment group and baseline score as covariates and a random intercept for trial.
- Model 2 (abbreviated total scores): Outcome is a simplified total score created after removing redundant items (as per relevant psychometric method) treatment group and corresponding baseline abbreviated score included as a covariates and a random intercept for trial.
    - Model 2a – factor analysis abbreviated total score
    - Model 2b – IRT abbreviated total score
    - Model 2c – NA abbreviated total score
- Model 3 (weighted psychometric scores): Outcome is a psychometrically derived score (e.g., factor score, IRT expected score, or net score) treatment group and corresponding baseline abbreviated score included as a covariates and a random intercept for trial.
    - Model 3a – factor analytic weighted scores
    - Model 3b – IRT weighted scores
    - Model 3c – NA weighted scores (net score)
- Model 4: Model 1 additionally adjusting for participant-level covariates (e.g. sex, age [where available])
- Model 5a-c: Model 2a-c additionally adjusting for the same participant-level covariates
- Model 6a-c: Model 3a-c additionally adjusting for the same participant-level covariates

Effect estimates from each model will be transformed into Cohen's d to facilitate comparison across trials and scoring approaches. The primary effect of interest is the difference between SMDs derived from those derived from conventional scoring and psychometrically-informed outcomes (i.e. Model 1 vs Models 2a-c vs Model 3a-c, and Model 4 vs 5a-c vs Model 6a-c).

Trial-level analyses

In addition to pooled IPD analyses, treatment effects will be estimated separately within each trial using both conventional and psychometrically-informed outcome scores. For each trial, change from baseline to outcome will be modelled adjusting for baseline severity, and trial-specific SMDs will be calculated for both scoring approaches. The difference between the psychometrically informed SMD and the conventional SMD will be computed for each trial and treated as the trial-level effect of interest. Potential moderators (e.g. sex) may be explored descriptively or within mixed models where data permits.

Secondary analyses: Response

Secondary analyses will explore whether psychometric refinement meaningfully alters estimates binary outcomes such as response. Response will be defined using established PANSS or BPRS definition.[42–44] Because binary outcomes may be sensitive to how thresholds map onto the underlying scale, these analyses will be treated as exploratory. The change in number of participants between psychometrically informed and original data will then be analysed using multilevel logistic regression models. Models will be fitted with and without adjustment for participant-level covariates.

Analyses will proceed in three steps:

1. Estimate baseline to outcome change in the proportion meeting response criteria within each arm, using both conventional and psychometrically-informed scoring.

2. Estimate treatment and comparator differences in response using conventional scoring and psychometrically-informed scoring in separate models.

3. Quantify the % difference between these two estimates (psychometrically-informed vs conventional) as the key comparison for secondary outcomes.

Exploratory subgroup and sensitivity analyses

Where feasible and supported by data availability, we will explore whether psychometric models and treatment effects vary by sex. Measurement invariance across sex will be assessed within the optimal psychometric models using established procedures. Where invariance is violated, group-specific psychometric models will be estimated, and primary and secondary analyses repeated using sex-specific scores.

To further examine the robustness of findings to potentially important sources of between-trial and clinical heterogeneity, we will attempt to conduct additional exploratory sensitivity analyses:

- Endpoint timing (week 4 vs week 6): Given that acute antipsychotic trials commonly report outcomes around week 6, but earlier endpoints (e.g., week 4) may yield different conclusions.[22] Therefore, we will, where endpoint timing can be harmonised and sample size allows, repeat analyses stratifying trials by the closest available acute endpoint (week 4 vs week 6).

- Stage of illness: Where sufficiently detailed and harmonisable information is available, we will explore whether the best-fitting psychometric model(s), invariance results, and estimated treatment effects differ according to patients' schizophrenia history (e.g., first-episode/early-phase versus multi-episode/chronic presentations, prior treatment exposure, or other trial-reported indicators of illness course).

These exploratory analyses are motivated by the clinical and methodological need to understand whether modelling choices or trial-design and population differences (such as endpoint timing or illness course) could materially influence inferences about treatment effects. All subgroup and sensitivity analyses will be clearly labelled exploratory, and interpretation will prioritise effect estimates and uncertainty, with transparent reporting of any constraints arising from limited sample sizes, inconsistent variable availability, or restricted harmonisability across trials, consistent with STROBE reporting principles.

Software and reproducibility

All analyses will be conducted using reproducible statistical workflows in R and RStudio, with additional software used for meta-analytic synthesis where appropriate. Analysis scripts and documentation will be shared in open repositories where permitted, in line with FAIR principles and Open Science guidelines.

Patient and Public involvement (PPI)

Patient and public involvement will be integrated to ensure that outcome interpretations and dissemination plans reflect the patient's priorities and are communicated in accessible language. Therefore, we will establish a panel including people with lived experience of psychosis/schizophrenia recruited through a clinic network in Ireland. The panel will meet 3–4 times for approximately 60 minutes via online sessions supported by brief material. Participants will be remunerated as per institutional policy. PPI provided input will include structured discussions and written feedback in (i) plain-language summary, (ii) framing of "clinically meaningful change" for PANSS/BPRS on an individual level, (iii) planned subgroup analysis reporting and equity implications, and (iv) dissemination formats. Results will be reported according to the GRIPP2 short form checklist.[45]

Discussion

Advanced psychometric methods are widely advocated in the methodological literature as tools to improve measurement quality in clinical research.[46,47] Yet, it remains uncertain whether they change what matters most – the estimated efficacy of interventions. This protocol is designed to test that practical question in schizophrenia research by quantifying whether psychometrically-informed scoring substantially alters antipsychotic treatment effect estimates.

Doyle et al. proposed,[19] and Byrne and colleagues subsequently demonstrated[13] that applying psychometric refinements can yield detectable differences in estimated antidepressant treatment effects, with the most consistent changes observed for factor-analytic approaches, while IRT- and network-based variants often produced little added value or inconsistent shifts in effects. The present project directly builds on this framework by translating it into schizophrenia trials, where measurement challenges and modest effects are long-standing concerns. Meta-analytic evidence from placebo-controlled antipsychotic drug trials have shown limited improvement across decades, suggesting that increasing placebo response is one plausible contributor to diminished effect sizes.[17] If placebo response grows (or baseline severity declines), detecting drug effects becomes more difficult, even when drugs remain efficacious, making it especially important that outcome measurement is as informative and low-noise as possible. Simultaneously, the interpretability of outcomes depends on the measurement properties of the instrument. A recent synthesis by Geck and colleagues concluded that the PANSS shows adequate performance regarding responsiveness and reliability, but has notable shortcomings in content validity and structural validity.[6] These limitations can weaken confidence in what total and subscale scores represent and how changes should be interpreted.

In this context, psychometric re-modelling using pooled IPD may help clarify which symptom structures replicate reliably, which items contribute meaningfully, and whether abbreviated scales will capture the same responsiveness to treatment or whether weighted scoring yields more accurate estimates of change. Importantly, we do not claim that psychometric refinement necessarily produces clinically meaningful change at the individual level. Rather, the planned analysis evaluates whether relatively modest differences in effect size (e.g., ~10–15% as observed in prior IPD psychometric work) are detectable and may improve the precision, interpretability, or comparability of trial effect estimates. Yet, they can still matter for meta-analytic synthesis, trial power, and the likelihood of detecting true treatment differences under challenging conditions.

Limitations

Several limitations should be acknowledged. Our planned analysis relies on available trials within controlled-access repositories: Hence, data might vary regarding follow-up timing, inclusion criteria, and data completeness, which may influence both psychometric models and effect size estimates. Psychometric BPRS and PANSS modelling across heterogeneous trials may yield more than one statistically defensible solution. We will therefore prioritise transparency in model comparison and report repository-specific results. Psychometric refinement may improve measurement in some contexts, but not in others. Data user agreements with Vivli and YODA Project may limit sharing of IPD and restrict reproducibility to code/scripts and permitted metadata, rather than full data transparency.

Ethics and consent

Ethical approval for the planned study was granted for the data collection by the Royal College of Surgeons in Ireland Research Ethics Committee (HREC Reference Number: REC202506033). Consent was not required for this study because no data were collected or analysed for this manuscript. The planned analyses will use secondary data, and consent for the use of participants' data in future research beyond the original study was obtained by the respective study sponsors.

Generative AI statement

The author(s) declared that Generative AI was used in the creation of this manuscript. During the preparation of this work authors used ChatGPT (GPT-4o mini, OpenAI) to refine the draft wording as well as a free online version of DeepL Write in order to cross-check the English phrasing and grammar. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Data availability

The data management plan and documentation describing the trials included in this study are openly available through Zenodo (v1.0; DOI: https://doi.org/10.5281/zenodo.19557064) and the Open Science Framework (v1.0; https://osf.io/xf6ac/files/6nb8g). Access to the underlying study data was obtained via Vivli (www.vivli.org) and the YODA Project (https://yoda.yale.edu/). The IPD used are not openly shareable because they comprise anonymised personal data and are subject to legal, ethical, and contractual restrictions, including GDPR requirements, consent restrictions, data use agreements, and sponsor or platform terms. Access to these data may be requested through Vivli and the YODA Project, subject to their respective governance and access requirements. Openly shareable materials will be provided via Zenodo and OSF under an open licence. These materials include non-disclosive derived outputs, such as aggregate trial-level estimates, psychometric parameters where permitted, model fit indices, analysis code, documentation, and synthetic or example data. These resources provide transparency regarding the trial selection and the analyses underlying the findings reported in this article.

Acknowledgment

This study, carried out under YODA Project # 2025-0836, will use data obtained from the Yale University Open Data Access Project, which has an agreement with JANSSEN RESEARCH & DEVELOPMENT, L.L.C. The interpretation and reporting of research using this data are solely the responsibility of the authors and does not necessarily represent the official views of the Yale University Open Data Access Project or JANSSEN RESEARCH & DEVELOPMENT, L.L.C. The original proposal can be found: https://yoda.yale.edu/data-request/2025-0836/. This publication is based on research who will be using data from data contributors [Abbott, AbbVie, Janssen, Johnson & Johnson, Lilly, Otsuka Pharmaceuticals] that has been made available through Vivli, Inc. Vivli has not contributed to or approved, and [Abbott, AbbVie, Janssen, Johnson & Johnson, Lilly, Otsuka Pharmaceuticals] are not in any way responsible for, the contents of this publication.

References

1. Kay SR, Fiszbein A, Opler LA: **The Positive and Negative Syndrome Scale (PANSS) for Schizophrenia.** *Schizophr Bull.* 1987 Jan 1; **13**(2): 261–276.
Publisher Full Text

2. Overall JE, Gorham DR: **The Brief Psychiatric Rating Scale.** *Psychol Rep.* 1962 Jun; **10**(3): 799–812.
Publisher Full Text

3. European Medicines Agency: *Guideline on clinical investigation of medicinal products, including depot preparations in the treatment of schizophrenia – Revision 1.* European Medicines Agency; 2012. Report EMA/CHMP/40072/2010 Rev. 1.
Reference Source

4. Abplanalp SJ, Green MF: **Symptom Structure in Schizophrenia: Implications of Latent Variable Modeling vs Network Analysis.** *Schizophr Bull.* 2022 May 7; **48**(3): 538–543.
Publisher Full Text

5. Baandrup L, Allerup P, Nielsen MØ, *et al.*: **Rasch analysis of the PANSS negative subscale and exploration of negative symptom trajectories in first-episode schizophrenia – data from the OPTiMiSE trial.** *Psychiatry Res.* 2020 Jul 1; **289**: 112970.
PubMed Abstract | Publisher Full Text

6. Geck S, Roithmeier M, Bühner M, *et al.*: **COSMIN systematic review and meta-analysis of the measurement properties of the Positive and Negative Syndrome Scale (PANSS).** *eClinicalMedicine.* 2025 Apr; **82**: 103155.
PubMed Abstract | Publisher Full Text | Free Full Text

7. Shafer A: **Meta-analysis of the Brief Psychiatric Rating Scale factor structure.** *Psychol Assess.* 2005; **17**(3): 324–335.
PubMed Abstract | Publisher Full Text

8. Shafer A, Dazzi F: **Meta-analysis of the positive and Negative Syndrome Scale (PANSS) factor structure.** *J Psychiatr Res.* 2019 Aug; **115**: 113–120.
PubMed Abstract | Publisher Full Text

9. Eronen MI, Bringmann LF: **The Theory Crisis in Psychology: How to Move Forward.** *Perspect Psychol Sci.* 2021 Jul 1; **16**(4): 779–788.
Publisher Full Text

10. Mokkink LB, Herbelet S, Tuinman PR, *et al.*: **Content validity: judging the relevance, comprehensiveness, and comprehensibility of an outcome measurement instrument – a COSMIN perspective.** *J Clin Epidemiol.* 2025 Sep; **185**: 111879.
PubMed Abstract | Publisher Full Text

11. Byrne D, Boland F, Brannick S, *et al.*: **Applying advanced psychometric approaches yields differential randomized trial effect sizes: secondary analysis of individual participant data from antidepressant studies using the Hamilton rating scale for depression.** *J Clin Epidemiol.* 2025 Jul; **183**: 111762.
PubMed Abstract | Publisher Full Text

12. Byrne D, Boland F, Doyle F: **A protocol for assessing stability and invariance properties of two commonly used outcome measures in anxiolytic clinical trials.** *HRB Open Res.* 2025 Sep 8; **8**: 106.
Publisher Full Text

13. Byrne D, Doyle F, Brannick S, *et al.*: **Do complex psychometric analyses really matter? Comparing multiple approaches using individual participant data from antidepressant trials.** *Psychol Med.* 2025; **55**: e289.
PubMed Abstract | Publisher Full Text | Free Full Text

14. Byrne D, Doyle F, Brannick S, *et al.*: **Evaluating the psychometric structure of the Hamilton Rating Scale for Depression pre- and post-treatment in antidepressant randomised trials: Secondary analysis of 6843 individual participants from 20 trials.** *Psychiatry Res.* 2024 Sep; **339**: 116057.
PubMed Abstract | Publisher Full Text

15. Byrne D, Ghoshal A, Boland F, *et al.*: **An exploratory graphical analysis of the Montgomery-Åsberg Depression Rating Scale pre- and post-treatment using pooled antidepressant trial secondary data.** *J Affect Disord.* 2025 Jan; **368**: 584–590.
PubMed Abstract | Publisher Full Text

16. The Lancet: **Time for renewed optimism for schizophrenia?.** *The Lancet.* 2024 Jan; **403**(10422): 117.
PubMed Abstract | Publisher Full Text

17. Leucht S, Leucht C, Huhn M, *et al.*: **Sixty Years of Placebo-Controlled Antipsychotic Drug Trials in Acute Schizophrenia: Systematic Review, Bayesian Meta-Analysis, and Meta-Regression of Efficacy Predictors.** *Am J Psychiatry.* 2017 Oct 1; **174**(10): 927–942.
PubMed Abstract | Publisher Full Text

18. European Comission: **rEVealIng Drug Effects when applyiNg psychomeTrics In schizophreniA triaLs (EVIDENTIAL).** 2025.
Publisher Full Text | Reference Source

19. Doyle F, Byrne D, Carney RM, *et al.*: **The effects of advanced factor analysis approaches on outcomes in randomised trials for depression: protocol for secondary analysis of individual participant data.** *BJPsych Open.* 2023 Sep; **9**(5): e157.
PubMed Abstract | Publisher Full Text | Free Full Text

20. Von Elm E, Altman DG, Egger M, *et al.*: **The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for Reporting Observational Studies.** *Ann Intern Med.* 2007 Oct 16; **147**(8): 573–577.
Publisher Full Text

21. McCutcheon RA, Harrison PJ, Howes OD, *et al.*: **Data-Driven Taxonomy for Antipsychotic Medication: A New Classification System.** *Biol Psychiatry.* 2023 Oct; **94**(7): 561–568.
PubMed Abstract | Publisher Full Text | Free Full Text

22. Younis IR, Gopalakrishnan M, Mathis M, *et al.*: **Association of End Point Definition and Randomized Clinical Trial Duration in Clinical Trials of Schizophrenia Medications.** *JAMA Psychiatry.* 2020 Oct 1; **77**(10): 1064–1071.
PubMed Abstract | Publisher Full Text | Free Full Text

23. Baandrup L, Allerup P, Nielsen MØ, *et al.*: **Scalability of the Positive and Negative Syndrome Scale in first-episode schizophrenia assessed by Rasch models.** *Acta Psychiatr Scand.* 2022 Jul; **146**(1): 21–35.
PubMed Abstract | Publisher Full Text | Free Full Text

24. Overall JE: **The Brief Psychiatric Rating Scale in Psychopharmacology Research.** Pichot P, Olivier-Martin R, editors. *Modern Trends in Pharmacopsychiatry.* S. Karger AG; 1974 [cited 2026 Jan 14]; pp. 67–78.
Publisher Full Text

25. Bajraktarov S, Blazhevska Stoilkovska B, Russo M, *et al.*: **Factor structure of the brief psychiatric rating scale-expanded among outpatients with psychotic disorders in five Southeast European countries: evidence for five factors.** *Front Psychiatry.*

2023 Oct 25; (14): 1207577.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

26. Hofmann AB, Schmid HM, Jabat M, *et al.*: **Utility and validity of the Brief Psychiatric Rating Scale (BPRS) as a transdiagnostic scale.** *Psychiatry Res.* 2022 Aug; **314**: 114659.
PubMed Abstract | **Publisher Full Text**

27. Lubke GH, Luningham J: **Fitting latent variable mixture models.** *Behav Res Ther.* 2017 Nov; **98**: 91–102.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

28. Hayton JC, Allen DG, Scarpello V: **Factor retention decisions in exploratory factor analysis: a tutorial on parallel analysis.** *Organ Res Methods.* 2004; **7**(Number 2): 191–205.
Publisher Full Text

29. Costello AB, Osborne J: **Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis.**
Publisher Full Text

30. Burns A, Höfer S, Curry P, *et al.*: **Revisiting the dimensionality of the Hospital Anxiety and Depression Scale in an international sample of patients with ischaemic heart disease.** *J Psychosom Res.* 2014 Aug; **77**(2): 116–121.
PubMed Abstract | **Publisher Full Text**

31. Quilty LC, Robinson JJ, Rolland JP, *et al.*: **The structure of the Montgomery-Åsberg depression rating scale over the course of treatment for depression.** *Int J Methods Psychiatr Res.* 2013 Sep; **22**(3): 175–184.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

32. Korkmaz S, Goksuluk D, Zararsiz G: **MVN: An R Package for Assessing Multivariate Normality.** *R J.* 2014; **6**(2): 151.
Publisher Full Text

33. Farahdel B, Thapaliya B, Suresh P, *et al.*: **Confirmatory Factor Analysis on Mental Health Status using ABCD Cohort.** *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM).* 2021 [cited 2026 Jan 19]; p. 3540–7.
Publisher Full Text | **Reference Source**

34. van der Ark LA: *Getting started with Mokken Scale Analysis in R.* Tilburg: Van der Ark; 2011.

35. Reise SP, Waller NG: **Item Response Theory and Clinical Measurement.** *Annu Rev Clin Psychol.* 2009 Apr 1; **5**(1): 27–48.
Publisher Full Text

36. Meijer RR, Baneke JJ: **Analyzing Psychopathology Items: A Case for Nonparametric Item Response Theory Modeling.** *Psychol Methods.* 2004; **9**(3): 354–368.
PubMed Abstract | **Publisher Full Text**

37. Chalmers RP: **mirt: A Multidimensional Item Response Theory Package for the R Environment.** *J Stat Softw.* 2012; **48**(6).
Publisher Full Text

38. Borsboom D, Cramer AOJ: **Network Analysis: An Integrative Approach to the Structure of Psychopathology.** *Annu Rev Clin Psychol.* 2013 Mar 28; **9**(1): 91–121.
Publisher Full Text

39. Borsboom D: **A network theory of mental disorders.** *World Psychiatry.* 2017 Feb; **16**(1): 5–13.
Publisher Full Text

40. Christensen AP, Golino H: **Estimating the Stability of Psychological Dimensions via Bootstrap Exploratory Graph Analysis: A Monte Carlo Simulation and Tutorial.** *Psych.* 2021 Aug 27; **3**(3): 479–500.
Publisher Full Text

41. Dickinson LM: **Multilevel Modeling and Practice-Based Research.** *Ann Fam Med.* 2005 May 1; **3**(suppl_1): S52–S60.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

42. Leucht S: **Measurements of Response, Remission, and Recovery in Schizophrenia and Examples for Their Clinical Application.** *J Clin Psychiatry.* 2014 Feb; **75**(suppl 1): 8–14.
PubMed Abstract | **Publisher Full Text**

43. Pinna F, Bosia M, Cavallaro R, *et al.*: **Consensus five factor PANSS for evaluation of clinical remission: effects on functioning and cognitive performances.** *Schizophr Res Cogn.* 2014 Dec; **1**(4): 187–192.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

44. Correll CU, Janetzky W, Brieden A: **Predictors of symptomatic and wellbeing remission in real-world samples of patients living with schizophrenia treated with aripiprazole once-monthly by means of constrained confidence partitioning.** *Schizophr Res.* 2025 Aug; **282**: 105–113.
PubMed Abstract | **Publisher Full Text**

45. Staniszewska S, Brett J, Simera I, *et al.*: **GRIPP2 reporting checklists: tools to improve reporting of patient and public involvement in research.** *BMJ.* 2017 Aug 2; j3453.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

46. Dima AL: **Scale validation in applied health research: tutorial for a 6-step R-based psychometrics protocol.** *Health Psychol Behav Med.* 2018 Jan 1; **6**(1): 136–161.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

47. Chan EKH: **Standards and Guidelines for Validation Practices: Development and Evaluation of Measurement Instruments.** Zumbo BD, Chan EKH, editors. *Validity and Validation in Social, Behavioral, and Health Sciences.* Cham: Springer International Publishing; 2014 [cited 2026 Jan 19]; pp. 9–24. (Social Indicators Research Series).
Publisher Full Text

Open Peer Review

Current Peer Review Status: ? ✓ ?

Version 1

Reviewer Report 28 August 2026

https://doi.org/10.21956/openreseurope.25577.r78518

© **2026 Iasevoli F.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Felice Iasevoli
University School of Medicine of Naples Federico II, Naples, Italy

This is a thoughtful and potentially valuable protocol addressing an important methodological question: whether psychometrically informed scoring of PANSS and BPRS meaningfully changes estimated antipsychotic treatment effects. The planned use of large-scale individual participant data, together with factor analytic, item-response and network approaches, is a major strength. I also appreciate the attempt to move beyond purely psychometric model fit and ask whether alternative scoring approaches materially influence treatment-effect estimation. I consider this a strong and worthwhile project. My main concern is not the use of industry-sponsored data per se, but whether psychometric models derived from a highly selected RCT ecosystem will generalize beyond that context, and whether a larger treatment-effect estimate will be interpreted too readily as evidence of better measurement. The study would be substantially strengthened by clearer separation between model development and out-of-sample treatment-effect evaluation, ideally at the trial level, together with a more explicit definition of what constitutes psychometric improvement and a stronger justification for network-derived weighting. These issues appear addressable within the proposed dataset and would, in my view, materially improve both the methodological rigor and the interpretability of the eventual findings. I describe these concerns below in more details:

The available datasets appear to come almost exclusively from pharmaceutical industry-sponsored trials accessed through YODA and Vivli. I do not regard industry sponsorship itself as a fatal limitation, particularly because the present study re-analyses randomized IPD independently of the original sponsor conclusions. However, I think the more important issue is the measurement context represented by these trials. Large industry-sponsored schizophrenia RCTs are highly selected environments, often involving restrictive eligibility criteria, standardized rater training, run-in or stabilization procedures, specific baseline severity requirements, and a limited range of antipsychotic agents.

Psychometric models derived from such data may therefore characterize PANSS/BPRS performance under the conditions of regulatory or commercial RCTs rather than necessarily generalizing to schizophrenia populations or antipsychotic treatment studies more broadly. Could

the authors clarify the intended target population to which the final psychometric models are meant to generalize? I would also encourage sensitivity analyses examining whether model structure and treatment-effect changes differ according to sponsor, trial design characteristics, and, where possible, medication class. If investigator-initiated trials cannot practically be included, this limitation should be framed primarily in terms of transportability rather than sponsorship bias alone.

The protocol appropriately separates exploratory and confirmatory datasets for EFA/CFA development. However, the main scientific question is not simply whether a psychometric structure can be replicated, but whether a scoring rule derived from that structure changes treatment-effect estimation in data not used to construct the rule. Could the authors clarify whether treatment-effect comparisons will be performed in data that are independent from those used to derive and select the psychometric models? With an expected sample of approximately 31,000 participants across many trials, a stronger validation strategy seems feasible. In particular, I wonder whether trial-level cross-validation could be considered. For example, psychometric scoring models could be derived in all but one trial and then applied to the held-out trial, iterating across studies. This would directly test whether psychometric refinement alters drug–comparator separation in a previously unseen RCT, which seems closer to the intended practical application than participant-level random splitting alone.

A central conceptual issue is that a larger treatment effect is not necessarily equivalent to better psychometric measurement. For example, a reweighted PANSS score could yield a larger drug–placebo standardized mean difference because it places greater weight on symptoms that are particularly treatment-responsive, without necessarily providing a more valid or more comprehensive measure of overall schizophrenia symptom severity. I therefore think the protocol should prespecify more clearly what will count as successful psychometric improvement. Is the primary criterion larger drug–placebo separation, greater statistical precision, lower measurement error, reduced between-trial heterogeneity, improved construct validity, improved responsiveness, or some combination of these? The interpretation of the study will depend heavily on this distinction.

I would encourage the authors to avoid equating a larger SMD per se with superior measurement and to define in advance how statistical improvement, psychometric validity and clinical relevance will be distinguished. The factor analytic and IRT approaches have a relatively clear rationale for deriving weighted or latent severity scores. I am less certain about the conceptual basis for translating network-analytic quantities into a single weighted treatment-outcome score. In network models, node strength or centrality reflects conditional relations among symptoms and does not necessarily imply that a symptom is more reliable, more informative regarding overall severity, or more appropriate to weight more heavily in a treatment outcome. Could the authors provide a stronger theoretical and statistical justification for deriving “net scores” from network loadings/strength? In particular, why should network centrality be expected to provide valid item weights for estimating treatment effects? This seems important because the network approach is conceptually distinct from latent-variable models and should not automatically be interpreted as an alternative method for estimating a unidimensional severity score.

The primary participant-level models include a random intercept for trial. Given the expected number and heterogeneity of included RCTs, I wonder whether this is sufficient. Treatment effects are likely to vary across trials because of differences in population, baseline severity, duration,

formulation, medication, run-in procedures, placebo response and other design characteristics. I therefore think the analysis plan should clarify whether treatment-by-trial heterogeneity will be explicitly modelled, for example through random treatment slopes, or whether a two-stage IPD meta-analytic approach with trial-specific treatment effects and random-effects synthesis will be the principal inferential framework. This issue seems particularly relevant because the main quantity of interest is the difference between conventional and psychometrically informed treatment effects. It would be informative not only to estimate the average difference, but also to determine whether that difference is consistent across trials or highly context-dependent.

I agree with the concern already raised by another reviewer regarding the derivation of separate baseline and post-treatment measurement models. If the measurement structure or item parameters differ across time, baseline and outcome scores may not represent the same construct on the same metric. Could the authors therefore clarify whether longitudinal measurement invariance will be formally assessed before interpreting score differences or treatment effects based on psychometrically informed scoring? In my view, this is particularly important for any model in which baseline-adjusted outcome scores are interpreted as reflecting treatment-related change.

The protocol contains a large number of psychometric approaches, alternative scores, subgroup analyses, drug classifications, outcome definitions and sensitivity analyses. This richness is valuable, but it also creates considerable analytical flexibility. I would therefore encourage the authors to define a clear hierarchy of primary psychometric model; primary scoring comparison; primary treatment-effect contrast; confirmatory versus exploratory analyses. The protocol should also state how multiplicity and selective interpretation across CFA, IRT, network models, abbreviated scores, weighted scores and subgroup analyses will be handled.

Is the rationale for, and objectives of, the study clearly described?

Yes

Is the study design appropriate for the research question?

Yes

Are sufficient details of the methods provided to allow replication by others?

Yes

Are the datasets clearly presented in a useable and accessible format?

Not applicable

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 14 August 2026

https://doi.org/10.21956/openreseurope.25577.r78521

© 2026 Sun Y. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Yaoyao Sun
Peking University Sixth Hospital, Peking University Institute of Mental Health, Beijing, China

Thanks for giving me the opportunity to review the interesting manuscript. This protocol addresses an important and clinically relevant methodological question: whether psychometrically informed scoring of the PANSS and BPRS can meaningfully alter estimates of antipsychotic treatment effects in schizophrenia trials. The proposed use of pooled individual participant data and multiple psychometric approaches, including factor analysis, item response theory, and network analysis, is promising. I hope several issues can be clarified:

1. The concept of “psychometric optimization” remains insufficiently defined. I would suggest the authors to clarify exactly what is being optimized. For example, item selection, symptom structure, weighting, measurement precision, or treatment-effect estimation and what changes are expected. Would the proposed methods increase drug–placebo separation, reduce between-trial heterogeneity, improve precision, or reduce sensitivity to placebo response?

2. It is unclear why investigator-initiated trials (IITs) are not included. High-quality IITs may provide important evidence and improve the generalizability of the findings. Conversely, relying mainly on pharmaceutical industry-sponsored trials may introduce sponsorship-related bias in study design, participant selection, and treatment-effect estimation. The authors should justify this restriction and, where possible, include sponsor type as a moderator or sensitivity analysis.

3. The Introduction acknowledges that psychometric validity may vary across populations and contexts, but cultural differences are not considered in the Methods. The authors should consider geographical region, language version, or cultural background in measurement-invariance or sensitivity analyses, where sufficient data are available.

4. The protocol should more clearly separate model development from treatment-effect estimation. Developing psychometric models and estimating treatment effects in the same data may lead to overfitting and inflated differences from conventional scoring.

5. The treatment effect reflects change from baseline to follow-up, but the protocol appears to develop separate baseline and post-treatment measurement models. If the factor structure or item parameters differ across time points, the resulting scores may not be directly comparable. Longitudinal measurement invariance should therefore be assessed before interpreting score changes as treatment effects.

6. The protocol includes many psychometric models, scoring approaches, subgroup analyses, and sensitivity analyses. The authors should clearly identify the primary model and primary comparison and describe how multiplicity and selective interpretation will be addressed.

7. Clinical interpretability should be considered alongside statistical improvement. Even if psychometrically informed scoring produces a larger effect size, it remains uncertain whether the

difference is clinically meaningful. The authors should prespecify how the magnitude and clinical relevance of any observed change will be evaluated.

Is the rationale for, and objectives of, the study clearly described?
Yes

Is the study design appropriate for the research question?
Yes

Are sufficient details of the methods provided to allow replication by others?
Yes

Are the datasets clearly presented in a useable and accessible format?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: biological psychiatry, clinical psychiatry, individualized treatment, pharmacogenomics, genome-wide association analysis

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 11 August 2026

https://doi.org/10.21956/openreseurope.25577.r77697

© 2026 Emsley R. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Robin Emsley
Stellenbosch University, Stellenbosch, Western Cape, South Africa

I have conducted this review from a clinical researcher viewpoint rather than a statistical perspective.

The manuscript describes a protocol aiming to assess the best psychometric models for the PANSS and BPRS scales for symptom assessment and treatment response in schizophrenia. The topic is of high importance. There is an enormous literature on studies of symptom severity and treatment outcome in schizophrenia, with many uncertainties remaining. The planned study has been carefully composed and includes a large sample of individuals from well characterised treatment outcome RCTs.

1. It is important to put the history of schizophrenia of rating scales into context: The PANSS was derived from BPRS, with additional items from the Psychopathology Rating Schedule added to provide more comprehensive assessment, particularly negative symptoms. The

PANSS has become overwhelmingly the instrument of choice to assess symptom severity and changes over time, largely because of its relative clarity and ease of use in clinical settings. The 30 selected items were originally grouped into 3 subscales, positive, negative and general psychopathology, based largely on clinicians' judgement. However, subsequent factor analyses indicated many of these items loaded with different domains, and some items loaded inconsistently across studies. Numerous FA studies followed, mostly favouring a five-factor solution. It should be kept in mind that criteria to define factors are somewhat arbitrary and methodology varied across studies. Many of these studies propose specific domain groupings and exclusion of certain items, basing their findings on extensive analyses. Much of what is published may have addressed the aims of this study, at least in part. It would be important for the present study to carefully review the current literature and define what is established as well as the uncertainties, and to clarify how this study will advance our understanding of the topic, rather than this being yet another analysis of the PANSS, just with a different statistical approach.

2. Multiple demographic and clinical variables will need to be carefully considered in the analyses. Also, medication dose and degree of adherence may significantly influence outcomes.
3. The studies that are planned for inclusion will overwhelmingly involve participants treated with risperidone of its active metabolite, paliperidone. While the advantage here is that it reduces possible confounding effects of differential efficacy effects of antipsychotics, it limits generalisability of these findings to people treated with other antipsychotics. So, for example, risperidone and paliperidone are more likely than some of the other antipsychotics to produce extrapyramidal symptoms that have been shown to effect some PANSS item scores.
4. Another consideration is that many of the RCTs (particularly those with long acting injectable formulations) have an initial run in treatment period to stabilise participants before baseline assessment. Analyses of treatment response in these studies will show much less effect than those assessing acutely psychotic individuals at baseline.
5. One of the major problems with PANSS is that scoring has been inconsistently applied regarding absence of symptoms. On the item scale of 1 to 7 a score of 1 indicates absence of a symptom, giving an absolute minimum total score of 30. Some studies correct for this in the analyses, while many don't. This has a particularly significant effect on outcomes that are reported as percentage reductions.
6. Minor comments:
    1. Paragraph 3 in Background: There are several sentences where words or phrases are omitted.
    2. Methods: Top of p 4: add *selected* before randomised controlled trials.

Is the rationale for, and objectives of, the study clearly described?

Yes

Is the study design appropriate for the research question?

Yes

Are sufficient details of the methods provided to allow replication by others?

Yes

Are the datasets clearly presented in a useable and accessible format?

Not applicable

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Clinical and treatment aspects of schizophrenia

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.