European Neuropsychopharmacology 94 (2025) 41–50

Contents lists available at ScienceDirect

European Neuropsychopharmacology

journal homepage: www.sciencedirect.com/journal/european-neuropsychopharmacology

COSMIN systematic review and meta-analysis of the measurement properties of the PANSS-6

Simon Geck [a], Maximilian Roithmeier [a], Markus Bühner [b], Sophia Wehr [a], Lucia Weigel [a], Josef Priller [a,c,d], John M. Davis [e], Stefan Leucht [a,f,*]

[a] Department of Psychiatry and Psychotherapy, School of Medicine and Health, Technical University of Munich, Klinikum rechts der Isar, Ismaningerstrasse 22, 81675, Munich, Germany
[b] Department of Psychology, Ludwig Maximilian University of Munich, Leopoldstr. 13, 80802 Munich, Germany
[c] Department of Psychiatry and Psychotherapy, School of Medicine and Health, Technical University of Munich, and German Center for Mental Health (DZPG), Klinikum rechts der Isar, Ismaninger Str. 22, 81675 Munich, Germany
[d] Neuropsychiatry and Laboratory of Molecular Psychiatry, Charité – Universitätsmedizin Berlin and DZNE, Berlin, Germany; Univeristy of Edinburgh and UK DRI, Edinburgh, UK
[e] Psychiatric Institute, University of Illinois at Chicago (mc 912), 1601W. Taylor St., Chicago, IL 60612, USA
[f] Technical University of Munich, School of Medicine and Health, Department of Psychiatry and Psychotherapy, Klinikum rechts der Isar, Ismaningerstrasse 22, 81675 Munich, Germany

ARTICLE INFO

Keywords:
Rating scale
Positive and negative syndrome scale
Meta-analysis
Psychometric properties
Psychosis
Antipsychotic drugs

ABSTRACT

The Positive and Negative Syndrome Scale (PANSS-30) is the standard instrument for assessing symptoms of schizophrenia and related psychotic disorders. However, its long administration time and structural issues have prompted the development of shorter versions. The PANSS-6, derived through Item Response Theory and Rasch analysis of the PANSS-8, emerged as a potential alternative. Comprising three positive and three negative symptom items, the PANSS-6 offers a more feasible assessment tool. However, its measurement properties have never been systematically reviewed.

We applied the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) guideline for systematic reviews and meta-analytical procedures to assess the psychometric properties of the PANSS-6. COSMIN comprises several steps: literature search, risk-of-bias assessments, assessing the updated criteria for good measurement properties, grading the quality of the evidence and feasibility aspects.

We included 13 publications. The PANSS-6 showed sufficient content validity, structural validity, measurement invariance, reliability, criterion validity, construct validity and responsiveness according to COSMIN. On some of them only a small body of evidence is currently available. For internal consistency, cross-cultural validity and measurement error there was not enough evidence for a definite rating. Its short administration time of only 15–20 mins renders feasibility good.

The PANSS-6 does not cover all schizophrenia symptoms but focuses on the core symptoms in favor of a feasible administration time. The evidence available on its measurement properties yields sufficient results for its purpose - the assessment of symptom severity and its change. According to COSMIN it can potentially be recommended for use.

1. Introduction

The 30-item Positive and Negative Syndrome Scale (PANSS) was published in 1987 (Kay et al., 1987) and has become the standard instrument for the assessment of schizophrenia symptoms. Later, multiple five-factor solutions emerged, which we addressed in our previous publication (Roithmeier et al., 2024). Nevertheless, both in its original three-subscale form and as a five-factor solution, the PANSS has shown shortcomings regarding its psychometric properties, e.g. method of development and poor structural validity (Geck and Roithmeier et al.,

* Corresponding author.
E-mail addresses: simon.geck@tum.de (S. Geck), maximilian.roithmeier@tum.de (M. Roithmeier), buehner@lmu.de (M. Bühner), sophia.wehr@tum.de (S. Wehr), lucia.weigel@tum.de (L. Weigel), davisjm@uic.edu (J.M. Davis), stefan.leucht@tum.de (S. Leucht).

https://doi.org/10.1016/j.euroneuro.2025.02.008
Received 28 October 2024; Received in revised form 12 February 2025; Accepted 14 February 2025
Available online 9 March 2025
0924-977X/© 2025 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

2024; Roithmeier et al., 2024). Its long assessment time of 45–60 min renders it relatively unpractical for routine use (Geck and Roithmeier et al., 2024; Kay et al., 1987), but even in research this long application time is a challenge. Applying item-response-theory (IRT)-analysis Santor et al. proposed a 14 item PANSS (Lindenmayer, 2017; Santor et al., 2007). (Ostergaard et al., 2016) assessed the scalability of the PANSS-30, PANSS-14 and a PANSS-8 following the remission criteria of (Andreasen et al., 2005) and found none of them to be scalable. This might be because some items do not provide unique information (Item heterogeneity) regarding the overall schizophrenia construct and due to person heterogeneity. One reason for person heterogeneity can be that some people answer the questions using a different response style. However, scalability is considered a statistical prerequisite for forming a total score as a measure of overall severity and to calculate reliability estimates (Bech, 2012; Lindenmayer, 2017). When (Ostergaard et al., 2016) omitted two items of the PANSS-8, the remaining six were scalable. This PANSS-6 comprises 3 positive symptom items (delusions, conceptual disorganization, hallucinatory behavior) and 3 negative symptom items (blunted affect, passive social withdrawal, lack of spontaneity and flow of conversation) (Ostergaard et al., 2016). Its shorter completion time (only 15–20 mins for a PANSS-6 assessment obtained with the Simplified-Negative-and-Positive Symptoms-Interview (SNAPSI), a clinical interview guide (Ostergaard et al., 2017) makes it a more feasible alternative to the PANSS-30. However, the psychometric properties of the PANSS-6 have to date never been assessed by a systematic review and meta-analysis. We fill this gap by applying the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) guideline (Mokkink et al., 2018; Prinsen et al., 2018; Terwee et al., 2018).

2. Methods

We follow the 2018 COSMIN guideline for systematic reviews of patient-reported outcome-measures which can also be applied to clinician-reported-outcome-measures (CLINROM) (Mokkink et al., 2018; Prinsen et al., 2018; Terwee et al., 2018). There are several steps: literature search, quality assessment of the individual studies with the COSMIN *Risk-of-Bias checklist*, applying *updated criteria for good measurement properties*, summarizing the evidence, grading the overall quality of evidence with a modified Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach, formulating a recommendation and describing feasibility aspects (Mokkink et al., 2018; Prinsen et al., 2018; Terwee et al., 2018). An overview is presented in Fig. 1.

All steps of the COSMIN methodology were conducted independently by two reviewers (SG and MR). Any disagreements were resolved by consensus, with the involvement of a third reviewer (SL, MB) if necessary.

COSMIN refers to individual assessments (not publications) as "studies". We follow this terminology. Therefore, the number of studies

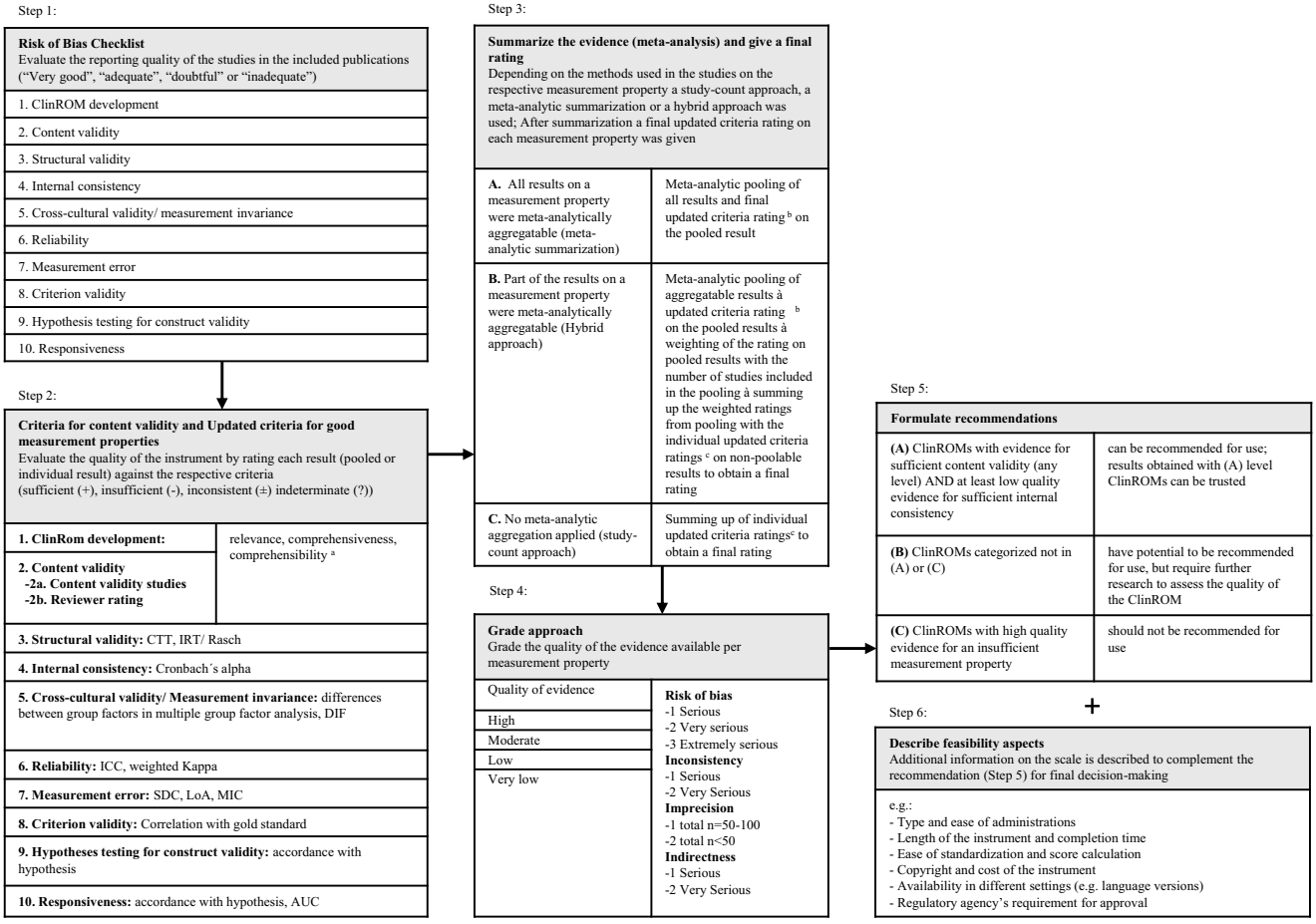


Fig. 1. Overview of the COSMIN methodology including adaptations
Refer to the methods part and appendix of our publication for further information. For the unmodified, full version of the COSMIN methodology, refer to the COSMIN manual. Step 1: Risk of Bias Checklist. Step 2: *Criteria for content validity* and *Updated criteria* for good measurement properties. [a] For the detailed criteria refer to the COSMIN manual for content validity. Step 3: Summarize the evidence and give a final rating. [b] The same criteria like in Step 2 apply. [c] These are the individual ratings on each result obtained in Step 2. Step 4: Grade approach. Step 5: Formulate recommendations. Step 6: Describe feasibility aspects.
SEE PDF.

differs from the number of publications.

A protocol with more details was registered on the Open-Science-Foundation website (https://doi.org/10.17605/OSF.IO/5EGMD; compare Appendix 1) (Roithmeier M, 2024).

2.1. Search strategy and selection criteria

Two reviewers (SG and MR) independently searched PubMed and EMBASE from inception to 01.10.2024 and without language or country restrictions with the term "six-item Positive and Negative Syndrome Scale" or "PANSS-6″. We screened the results and reference lists of included publications for evaluation studies of the PANSS-6 in which at least 80 % of study participants had a psychotic disorder. Following COSMIN, only full-text articles were included (Mokkink et al., 2018; Prinsen et al., 2018; Terwee et al., 2018).

2.2. Assessing the risk of bias (step 1 in Fig. 1)

The COSMIN risk of bias checklist is a tool for assessing *how well* the following 10 psychometric domains (called "boxes" by COSMIN) were assessed: (1) ClinROM development, (2) content validity, (3) structural validity, (4) internal consistency, (5) cross-cultural validity/ measurement invariance, (6) reliability, (7) measurement error, (8) criterion validity, (9) hypothesis testing for construct validity, (10) responsiveness (Mokkink et al., 2018; Prinsen et al., 2018; Terwee et al., 2018). Each box consists of two to eight items rated as either "very good", "adequate", "doubtful", "inadequate" or "not applicable". A *worst score counts* principle is applied for each domain. Only those domains which were actually evaluated in a publication are assessed.

2.3. Assessing the criteria for content validity and the updated criteria for good measurement properties (step 2 in Fig. 1)

In step 2 *the results* of each study are evaluated. a) *Content validity* is addressed by domains 1 and 2. The judgement involves the original ClinROM development study, content validity studies and a reviewer rating of the scale (Fig. 1, step 2, domains 1 and 2; and Appendix 2) b) The *Updated criteria* concern domains 3–10. For assessing domains 9 "construct validity" and 10 "responsiveness", COSMIN requires a formulation of hypotheses (Mokkink et al., 2018; Prinsen et al., 2018; Terwee et al., 2018) which are presented in Appendix 3. In cases where statistics (e.g. Cronbach's alpha) were calculated in the same sample at different time points we calculated averages to prevent a disproportionate influence of them. For details about the assessment of all other domains suggested by COSMIN, please see pp.28–29 of the COSMIN manual (Mokkink et al., 2018; Prinsen et al., 2018; Terwee et al., 2018).

A study's evidence for each domain is rated as either "sufficient" (+) or "insufficient" (-), or "indeterminate" rating (?) if neither "sufficient" nor "insufficient" ratings are applicable. Only for content validity (domains 1 and 2), a fourth option of "inconsistent" (±) is available.

2.4. Statistical method and summarizing the evidence (Step 3 in Fig. 1)

Results on the measurement properties *internal consistency, interrater-reliability, criterion validity* and *construct validity* were summarized with standard random-effects meta-analyses using Comprehensive Meta-Analysis (CMA) Version 2 (Borenstein et al., 2009) (Fig. 1, Step 3, Option A). We measured heterogeneity with the I^2-statistic and a chi-square test. For *responsiveness* it was not possible to aggregate *all* studies. We, therefore, used a hybrid of the meta-analytic approach and the simple study count suggested by COSMIN. I.e. the meta-analytic result was weighted by the number of included studies, and was then summed up with ratings of studies that could not be aggregated (Fig. 1, Step 3, Option B). Results on structural validity, measurement invariance and measurement error could not be pooled at all. Therefore, COSMIN's study count approach applied (Fig. 1, Step 3, Option C). If at least 65 % of the ratings agree on the in- or sufficiency of the results on one measurement property, the corresponding final rating ("sufficient" or "insufficient") is given. If <65 % of results are consistent across the single ratings, the overall rating of the respective measurement property is considered as "inconsistent". "Indeterminate" ratings of single studies are not considered regarding inconsistencies.

2.5. Grading the quality of evidence (Step 4 in Fig. 1)

A Grading of Recommendations Assessment, Development, and Evaluation (GRADE) approach modified by COSMIN was used to grade the confidence in the reliability of the results (GRADE, 2013; Mokkink et al., 2018; Prinsen et al., 2018; Terwee et al., 2018). This procedure was applied to each measurement property separately. The four factors that determine GRADE were: (1) risk of bias, (2) inconsistency (of the results of the studies), (3) imprecision (total sample size) and (4) indirectness (evidence from situations which do not fully match with the review question).

Initially, evidence was assumed to be of high quality. If the criteria of one or more factors were not fulfilled, the evidence was downgraded by one to three levels. Ultimately, the confidence in the evidence was graded as high, moderate, low, or very low (see Table 1).

The grading of the quality of evidence was only done for results on measurement properties that received a "sufficient" or "insufficient"

Table 1
GRADE approach.

a) Quality of evidence definitions	
High	We are very confident that the true measurement property lies close to that of the estimate of the measurement property
Moderate	We are moderately confident in the measurement property estimate: the true measurement property is likely to be close to the estimate of the measurement property, but there is a possibility that it is substantially different
Low	Our confidence in the measurement property estimate is limited: the true measurement property may be substantially different from the estimate of the measurement property
Very low	We have very little confidence in the measurement property estimate: the true measurement property is likely to be substantially different from the estimate of the measurement property
b) Instructions on downgrading	
Risk of bias	
No	There are multiple studies of at least adequate quality, or there is one study of very good quality available
−1 Serious	There are multiple studies of doubtful quality available, or there is only one study of adequate quality
−2 Very serious	There are multiple studies of inadequate quality, or there is only one study of doubtful quality available
−3 Extremely serious	There is only one study of inadequate quality available
Inconsistency	
No	Results could be meta-analytically pooled or ≥75 % of the studies had consistent results
−1 Serious	<75 % of studies had consistent results
−2 Very serious	<70 % and ≥65 % had consistent results If <65 % had consistent results GRADE will not be performed because there is not enough evidence
Imprecision[a]	As recommended by COSMIN due to the high total sample sizes no downgrading for imprecision was done
Indirectness	As recommended by COSMIN due to our inclusion criteria of ≥80 % population with a psychotic disorder no downgrading for indirectness was done

[a]Downgrading for Imprecision is not applicable for measurement properties in which the sample size is already evaluated in the "Risk of bias" step (studies on content validity, structural validity and cross-cultural validity/measurement invariance).

rating in the "updated criteria for good measurement properties" step. "Indeterminate" results were not graded.

2.6. Recommendations (Step 5 in Fig. 1)

According to COSMIN, an overall judgement of (A) "can be recommended for use and results obtained with the scale can be trusted", (B) "potential for recommendation but needs further research into the measurement properties", or (C) "should not be recommended for use" was made (Mokkink et al., 2018; Prinsen et al., 2018; Terwee et al., 2018).

2.7. Feasibility (Step 6 in Fig. 1)

Feasibility aspects of the PANSS-6 were summarized as recommended by COSMIN.

3. Results

3.1. Literature search

The search yielded 48/ 17 results in EMBASE/ PubMed. After title-abstract-screening, 18 publications were selected for full-text-screening and 13 were then identified as eligible for this review. See Appendix 4 for a depiction of this process and Appendix 5 for the general characteristics of included articles populations.

3.2. Assessing the risk of bias (Table 2, left side)

Content validity (= ClinROM development and content validity)

ClinROM development: By definition, ClinROM development is not considered a measurement property. However, it is considered in the assessment of content validity. It involves evaluating whether design requirements were met, and whether comprehensibility and comprehensiveness were adequately addressed in pilot testing (Mokkink et al., 2018; Prinsen et al., 2018; Terwee et al., 2018; Weigel et al., 2023). The PANSS-6 was not developed from scratch but derived from the PANSS-30 (Ostergaard et al., 2016). Therefore, no development study for the PANSS-6 by COSMINs terms exists. Unfortunately, the PANSS-6 can also not 'lend' this aspect from the development process of the PANSS, because the PANSS-30 has also not been appropriately developed. In essence the PANSS-30 was a merger of the BPRS and the unvalidated Psychopathology Rating Schedule (PRS) (Singh and Kay, 1975). Moreover, aspects such as the involvement of clinicians and patients have never been done (Geck and Roithmeier et al., 2024).

3.2.1. Content validity

Content validity studies refer to further studies asking patients and professionals about the relevance, comprehensiveness or comprehensibility of an existing ClinROM. Such studies can be performed by the developers of the ClinROM or by other researchers (Mokkink et al., 2018; Prinsen et al., 2018; Terwee et al., 2018; Weigel et al., 2023). No such studies could be identified.

3.2.2. Internal structure

Structural Validity: Structural validity measures the degree to which the scores of an instrument are an adequate reflection of the dimensionality of the construct to be measured (Mokkink et al., 2018; Prinsen et al., 2018; Terwee et al., 2018). Of 18 included studies on structural validity (Baandrup et al., 2022; Li et al., 2021; Lin et al., 2018; Ostergaard et al., 2018a, b; Ostergaard et al., 2016), 1 conducted a confirmatory factor analysis (Li et al., 2021) and 17 assessed the scalability of the PANSS-6 (Baandrup et al., 2022; Lin et al., 2018; Ostergaard et al., 2018a, b; Ostergaard et al., 2016). All 18 studies (Baandrup et al., 2022; Li et al., 2021; Lin et al., 2018; Ostergaard et al., 2018a, b; Ostergaard et al., 2016) received "very good" ratings.

Internal consistency: Internal consistency measures the interrelatedness among items, with Cronbach's Alpha (α) being the most common statistic. Three studies (Li et al., 2021; Lin et al., 2018; Pagsberg et al., 2022) were identified. All were rated "inadequate", because internal consistency statistics should not be calculated for scores of scales that are not unidimensional or consist of subscores (Mokkink et al., 2018; Prinsen et al., 2018; Terwee et al., 2018), but (Kolbaek et al., 2021b; Nielsen et al., 2022) have described a positive and a negative subscale of PANSS-6, and (Li et al., 2021) found a very good fit for a two-factor solution. (see *Updated criteria* section below).

Cross-cultural validity/Measurement invariance: COSMIN differentiates between the "cross-cultural validity" and "measurement invariance" depending on the used methodology. *Measurement invariance* specifically refers to whether respondents from different groups with the same latent trait level (allowing for group differences) respond similarly to a particular item (Mokkink et al., 2018; Prinsen et al., 2018; Terwee et al., 2018). In our case the latent trait is schizophrenia symptom severity. Of five studies assessing *measurement invariance,* three (Baandrup et al., 2022; Lin et al., 2018; Ostergaard et al., 2018a) were rated "doubtful", because it was not specified whether samples were similar for relevant characteristics except for the group variable (Mokkink et al., 2018; Prinsen et al., 2018; Terwee et al., 2018). The remaining two studies (Ostergaard et al., 2018b, 2016) were rated "inadequate" because they additionally did not meet COSMINs sample size requirements.

3.2.3. Remaining measurement properties

Reliability: Two studies (Kolbaek et al., 2018, 2021a) assessing the *interrater reliability* of the PANSS-6 were rated "adequate" as test conditions were not clearly reported to be similar for different administrations and as the weighting scheme used for statistical calculations was not described (Mokkink et al., 2018; Prinsen et al., 2018; Terwee et al., 2018). No studies on *test-retest reliability* were identified.

Measurement error: Measurement error refers to the systematic and random error of a patient's score that is not attributed to true changes in the construct to be measured (Mokkink et al., 2018; Prinsen et al., 2018; Terwee et al., 2018). One study (Baandrup et al., 2022) on *measurement error* was identified and only rated "adequate" because the participants were not stable between measurements.

Criterion validity: The PANSS-30 from which the PANSS-6 has been derived and which has demonstrated good construct validity (Geck and Roithmeier et al., 2024) was used as a gold-standard. Seven studies (Kolbaek et al., 2021b; Li et al., 2021; Lin et al., 2018; Nielsen et al., 2022; Ostergaard et al., 2018b, 2017; Pagsberg et al., 2022) were identified and used to assess *criterion validity*. All of them were rated as "very good".

Hypothesis testing for construct validity: Two publications (Kolbaek et al., 2021b; Nielsen et al., 2022) correlated PANSS-6 scores to PANSS-8 scores which are used in the remission criteria of (Andreasen et al., 2005). As there is only some evidence on the measurement properties of the PANSS-8 available, both studies received "doubtful" ratings. While the PANSS-6 was initially not described as a two-dimensional scale (Ostergaard et al., 2016), later authors described two subscales of three items each (items P1, P2, P3 on the positive subscale, items N1, N4 and N6 for the negative subscale) (Kolbaek et al., 2021b; Li et al., 2021; Nielsen et al., 2022). The four assessments of these subscales (Kolbaek et al., 2021b; Nielsen et al., 2022) were rated as "doubtful" (Geck and Roithmeier et al., 2024). The "doubtful" rating is primarily based on the insufficient structural validity of the original PANSS-30, whose 3-subscale structure has been refuted and thus also contradicts the original positive and negative subscales (Geck and Roithmeier et al., 2024).

Responsiveness: Responsiveness refers to the ability of an instrument to detect change over time (Mokkink et al., 2018; Prinsen et al., 2018; Terwee et al., 2018). Of 21 studies which assessed the ability of the PANSS-6 to detect symptom remission, and antipsychotic-placebo or antipsychotic-antipsychotic separation, 101 studies (Hieronymus et al.,

2021; Kolbaek et al., 2021b; Li et al., 2021; Lin et al., 2018; Nielsen et al., 2022; Ostergaard et al., 2018b, 2016, 2017) were rated "very good", 5 studies (Hieronymus et al., 2021; Li et al., 2021; Ostergaard et al., 2018a, b) "doubtful", and 6 studies (Hieronymus et al., 2021; Lin et al., 2018; Ostergaard et al., 2018a, b; Pagsberg et al., 2022) were rated "inadequate". The doubtful ratings were assigned because important characteristics of compared subgroups were not comprehensibly provided and the "inadequate" ratings were assigned because methods were used, that COSMIN does not deem suitable for the assessments in question (e.g. COSMIN requires the calculation of correlations, area under the curve (AUC) or sensitivity and specificity for the criterion approach in the assessment of responsiveness and the authors used e.g. effect size differences; please see COSMIN manual page 60–63 for details) (Mokkink et al., 2018; Prinsen et al., 2018; Terwee et al., 2018).

Assessing the criteria for content validity and the updated criteria for good measurement properties (Table 2, right side)

3.3. Assessing the criteria for content validity

The PANSS-6 was derived from the PANSS-30 (Ostergaard et al., 2016), thus no development study by COSMIN definitions was available, and there were no studies on content validity. Therefore, the rating on content validity is only based on the rating of the review team.

The *content validity* ratings were "sufficient" for *Relevance*, "indeterminate" for *Comprehensiveness* and "sufficient" for *Comprehensibility*. *Comprehensiveness was rated* "indeterminate" because COSMIN requires that a scale's items fully capture the construct being measured (Mokkink et al., 2018; Prinsen et al., 2018; Terwee et al., 2018). The PANSS-6, however, covers only four out of the five schizophrenia criteria in DSM-5 (Association, 2013; Ostergaard et al., 2016). The resulting overall rating for content validity was "sufficient".

3.4. Assessing the updated criteria for good measurement properties

Forest-plots of meta-analyses of *internal consistency, interrater-reliability, construct validity* and *responsiveness* are presented in Panel 1.

3.4.1. Internal structure

Structural validity: (Li et al., 2021) conducted a confirmatory factor analysis of the two-factor (positive and negative symptom) model of the PANSS-6. The Comparative Fit Index (CFI=0.99) and the Root-Mean-Square-Error-of-Approximation (RMSEA=0.04) surpassed the COSMIN criterion of CFI>0.95 or RMSEA<0.06 for good model fit leading to the rating "sufficient". 17 studies (Baandrup et al., 2022; Lin et al., 2018; Ostergaard et al., 2016, 2018a, 2018b) on scalability were rated "indeterminate" because none of them fulfilled *all* 4 criteria required by COSMIN ~~criteria~~ for Item Response Theory-/ Rasch-analyses for a "sufficient"/ "insufficient" rating („no violation of unidimensionality", „no violation of local independence", „no violation of monotonicity" and „adequate model fit"; compare COSMIN manual page 28) (Mokkink et al., 2018; Prinsen et al., 2018; Terwee et al., 2018).

Internal consistency: Three studies (Li et al., 2021; Lin et al., 2018; Pagsberg et al., 2022) resulted in a pooled Cronbach's α of 0.70 (rounded), just surpassing the COSMIN cutoff of α≥0.70. However, an "indeterminate" rating was given because Cronbach's alpha should only be calculated for unidimensional scales or subscales.

Measurement invariance: Five studies assessed the influence of the group variables age, gender, treatment and time by means of Differential Item Functioning (DIF) analyses. Four (Lin et al., 2018; Ostergaard et al., 2016, 2018a, 2018b) didn't detect any DIF which means that patients differing on the above-mentioned group variables (e.g., different sex) with the same latent symptom severity have an equal probability of item scoring (Baandrup et al., 2022). Those four were rated "sufficient". The fifth (Baandrup et al., 2022) found DIF between age and gender groups and was therefore rated "insufficient". However, the presence of external influencing and untested variables could not be clearly ruled out, making the determination whether DIF can be attributed to age or gender differences hard. The overall rating for *measurement invariance* was "sufficient".

3.4.2. Remaining measurement properties

Interrater reliability: One (Kolbaek et al., 2018) study reported an Intraclass Correlation Coefficient (ICC) and a second (Kolbaek et al., 2021a) a Gwet's agreement coefficient. The pooled *interrater reliability was* 0.77, surpassing the COSMIN cutoff of ICC≥0.70 (Mokkink et al., 2018; Prinsen et al., 2018; Terwee et al., 2018), resulting in a "sufficient" rating for *interrater reliability*.

Measurement error: One study (Baandrup et al., 2022) calculated a Standard Error of Measurement (SEM) of 0.52, but SEM is not a method COSMIN provides *updated criteria* for. The study was therefore rated "indeterminate". However, an SEM of 0.52 means that an individual's observed score on PANSS-6 should be expected to fluctuate around the "true score" by an average of 0.52 points due to error.

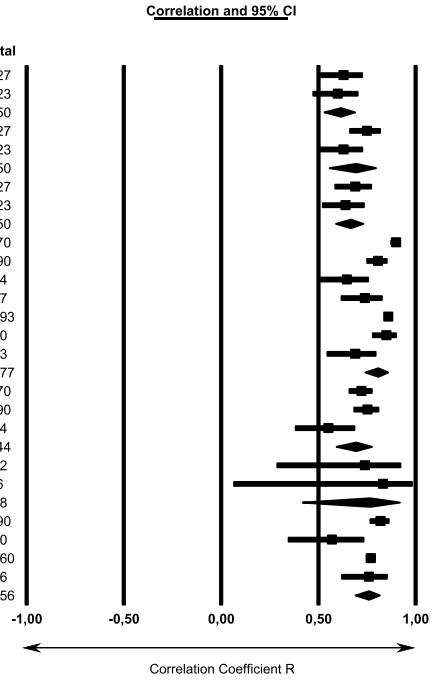
Criterion validity: Seven studies (Kolbaek et al., 2021b; Li et al., 2021; Lin et al., 2018; Nielsen et al., 2022; Ostergaard et al., 2018b, 2017; Pagsberg et al., 2022) correlated the PANSS-6 with the PANSS-30 and

Table 2
COSMIN risk of bias and criteria results.

Measurement property	Scale (No. of studies assessing the scales measurement property)	COSMIN risk of bias				COSMIN updated criteria ratings*			
		Very good	Adequate	Doubtful	Inadequate	+[a]	-[b]	?[c]	±[d]
ClinROM Development	PANSS-6 ($n = 0$)	0	0	0	0	0	0	0	0
Content validity	PANSS-6 (content validity studies; $n = 0$)	0	0	0	0	0	0	0	0
	PANSS-6 (content validity reviewer rating)	/	/	/	/	1	0	0	0
Structural validity	PANSS-6 ($n = 18$)	18	0	0	0	1	0	17	/
Internal consistency	PANSS-6 ($n = 3$)	0	0	0	3	0	0	3	/
Cross-cultural validity	PANSS-6 ($n = 0$)	0	0	0	0	0	0	0	/
Measurement invariante	PANSS-6 ($n = 5$)	0	0	3	2	4	1	0	/
Interrater reliability	PANSS-6 ($n = 2$)	0	2	0	0	2	0	0	/
Test-retest reliability	PANSS-6 ($n = 0$)	0	0	0	0	0	0	0	/
Measurement error	PANSS-6 ($n = 1$)	0	1	0	0	0	0	1	/
Criterion validity	PANSS-6 ($n = 7$)	5	0	2	0	4	2	0	/
Hypothesis testing for construct validity	PANSS-6 ($n = 2$)	0	0	0	2	2	0	0	/
	PANSS-6 positive subscale ($n = 2$)	0	0	2	0	2	0	0	/
	PANSS-6 negative subscale ($n = 2$)	0	0	2	0	2	0	0	/
Responsiveness	PANSS-6 ($n = 22$)	11	0	5	6	19	2	1	/

*ClinROM development rating resulted from *Criteria for content validity;* [a] "Sufficient"; [b] "Insufficient"; [c] "Indeterminate"; [d] "Inconsistent".

Study name	Comparison	Correlation	Lower limit	Upper limit	Z-Value	p-Value	Total
Kolbaek2021	ConV-PANSS6negativesubscale-PANSS30negativesubscale	0,630	0,512	0,725	8,256	0,000	127
Nielsen2022	ConV-PANSS6negativesubscale-PANSS30negativesubscale	0,600	0,473	0,702	7,593	0,000	123
		0,615	0,531	0,687	11,210	0,000	250
Kolbaek2021	ConV-PANSS6-PANSS8	0,750	0,662	0,817	10,834	0,000	127
Nielsen2022	ConV-PANSS6-PANSS8	0,630	0,510	0,726	8,122	0,000	123
		0,695	0,559	0,795	7,409	0,000	250
Kolbaek2021	ConV-PANSS6positivesubscale-PANSS30positivesubscale	0,690	0,586	0,771	9,442	0,000	127
Nielsen2022	ConV-PANSS6positivesubscale-PANSS30positivesubscale	0,640	0,522	0,734	8,305	0,000	123
		0,666	0,590	0,730	12,556	0,000	250
Lin2018	CritV-PANSS6	0,900	0,875	0,920	24,056	0,000	270
Li2021	CritV-PANSS6	0,807	0,751	0,852	15,293	0,000	190
Pagsberg2022	CritV-PANSS6	0,647	0,502	0,756	6,931	0,000	84
Kolbaek2021	CritV-PANSS6	0,740	0,619	0,827	8,176	0,000	77
Ostergaard2017	CritV-PANSS6	0,860	0,846	0,873	49,924	0,000	1493
Ostergaard2018	CritV-PANSS6	0,850	0,780	0,899	11,717	0,000	90
Nielsen2022	CritV-PANSS6	0,690	0,547	0,794	7,095	0,000	73
		0,808	0,742	0,858	13,352	0,000	2277
Lin2018	IC-PANSS6	0,722	0,659	0,775	14,899	0,000	270
Li2021	IC-PANSS6	0,753	0,684	0,809	13,399	0,000	190
Pagsberg2022	IC-PANSS6	0,552	0,383	0,685	5,591	0,000	84
		0,695	0,594	0,775	9,685	0,000	544
Kolbaek2018	IR-PANSS6	0,740	0,289	0,922	2,851	0,004	12
Kolbaek2021	IR-PANSS6	0,833	0,066	0,981	2,075	0,038	6
		0,767	0,419	0,918	3,507	0,000	18
Li2021	Resp-PANSS6changescore-PANSS30changescore	0,820	0,767	0,862	15,819	0,000	190
Kolbaek2021	Resp-PANSS6changescore-PANSS30changescore	0,570	0,347	0,732	4,439	0,000	50
Ostergaard2017	Resp-PANSS6changescore-PANSS30changescore	0,770	0,748	0,790	38,947	0,000	1460
Ostergaard2018	Resp-PANSS6changescore-PANSS30changescore	0,760	0,621	0,853	7,253	0,000	56
		0,760	0,691	0,815	13,339	0,000	1756



Panel 1. Results from meta-analysis

Standard random-effects meta-analyses were conducted using Comprehensive Meta-Analysis Version 2; Diamond: Represents the pooled correlation from the meta-analysis. Its width indicates the 95 % confidence interval (CI). Line and Cube: Each cube represents an individual study's correlation, with its size reflecting the study's weight in the analysis. The horizontal line shows the 95 % CI for that study.

Line and Cube: Each cube represents an individual study's correlation, with its size reflecting the study's weight in the analysis. The horizontal line shows the 95 % CI for that study. Tests for heterogeneity detected statistically significant heterogeneity in ConV-PANSS6-PANSS8, CritV-PANSS6, IC-PANSS6 and Resp-PANSS6changescore-PANSS30changescore. However, correlations were reasonably close together and all correlations for which heterogeneity was detected were >0.5. Thus, while being statistically significant, the heterogeneity was not clinically relevant and may in part be due to small sample sizes (Schönbrodt & Perugini, 2013). Abbreviations: ConV - Convergent validity; CritV - Criterion validity; IC - Internal consistency; Resp - Responsiveness.

SEE PDF/PPTX.

resulted in a "sufficient" rating, because the pooled correlation of ~~0.79~~ 0.81 surpassed COSMINS cutoff of ≥0.70.

Hypothesis testing for construct validity: Two studies (Kolbaek et al., 2021b; Nielsen et al., 2022) correlated the PANSS-6 with the PANSS-8 and the pooled correlation of 0.70 (rounded) confirmed *hypothesis 1* (correlation≥0.5; see Appendix 3). A pooled correlation of 0.67/ 0.62 between the positive/negative subscales of PANSS-6 and PANSS-30, both calculated from two studies each, confirmed *hypothesis 1* as well. Therefore, the construct validity of both PANSS-6 total and its subscales were rated "sufficient" overall.

Responsiveness: 22 studies covered different ways and aspects to assess *responsiveness*: 1. Remission detection, 2. Treatment efficacy and placebo differentiation and 3. Comparison of change scores. The studies of each of these three aspects are numbered for easier comprehensibility. Across the respective 9 publications different states of schizophrenia were covered. To rate responsiveness, we used the pre-formulated *hypotheses 2–8* (see Appendix 3). For an overview of all ratings compare Table 2.

Remission detection: 5 studies assessed the ability of the PANSS-6 to detect remission. The 8 PANSS items (PANSS-8) suggested by (Andreasen et al., 2005) and a score on each of them of ≤3 were operationalized as remission-criteria.

Studies 1–4/5 (Hieronymus et al., 2021; Li et al., 2021; Ostergaard et al., 2018a, b) reported sensitivity/specificity values of 1.00/0.86–0.99. Therefore, they confirmed our *hypothesis 5* (sensitivity/specificity ≥0.8) and resulted in 4 "sufficient" ratings.

Study 5/5 (Lin et al., 2018) presented an Area Under the Curve (AUC) of 0.82 regarding the detection of remission. It surpassed COSMINs criterion of AUC ≥0.70 and was thus rated "sufficient".

Treatment efficacy and placebo differentiation: 12 studies assessed the sensitivity of the PANSS-6 to antipsychotic "before versus after" efficacy and antipsychotic-placebo separation:

Studies 1–4/12 (Hieronymus et al., 2021; Lin et al., 2018; Nielsen et al., 2022; Pagsberg et al., 2022) reported baseline to post-baseline effects. Effect sizes ranging from 0.29 to 1.09 from antipsychotic trials. They therefore confirmed our hypothesis 2 (ES>0.2) and resulted in 3 "sufficient" ratings.

Studies 5–7/12 compared the intervention effect sizes calculated from PANSS-6 and PANSS-30 data. In two (Hieronymus et al., 2021; Lin et al., 2018) of them both scales' effect sizes differed ≤20 % and therefore confirmed our *hypothesis 4*. The effect size difference in the third study (Pagsberg et al., 2022) was >20 % and did not confirm *hypothesis 4*. Accordingly, 2 "sufficient" ratings and 1 "insufficient" rating were issued.

Study 8/12 (Li et al., 2021) reported sensitivity/specificity values of 0.77/0.84 to predict response defined as at least 50 % PANSS-30 total score reduction from baseline and could not confirm our *hypothesis 5* (sensitivity/specificity ≥0.8) and was therefore rated "insufficient".

Study 9/12 (Lin et al., 2018) presented an AUC=0.84 for response detection (defined as at least 40 % PANSS-30 total score reduction from baseline), surpassed the COSMIN criterion of AUC≥0.7 and was therefore rated "sufficient".

Study 10/12 (Ostergaard et al., 2018a) found that the PANSS-6 detected the same relevant differences as the PANSS-30

(Bonferroni-adjusted level=0.005) and therefore confirmed our *hypothesis 6*. The study was therefore rated "sufficient".

Study 11/12 (Ostergaard et al., 2016) calculated the Number-Needed-to-Treat (NNT) for remission to separate the efficacy of haloperidol/sertindole and placebo. As haloperidol is an established and well investigated antipsychotic and little data are available on sertindole, we focused on haloperidol (Huhn et al., 2019). The reported NNT was 9/6 after 6/8 weeks and therefore confirmed our *hypothesis 3* (NNT<10). The study was therefore rated "sufficient".

For Study 12/12 (Ostergaard et al., 2018b) which reported the speed of change measured on PANSS-6 no hypothesis could be formulated and it was therefore rated "indeterminate".

The respective authors of these 12 studies concluded that the PANSS-6 is suitable for detecting treatment-effects and antipsychotic-placebo separation.

Comparison of change scores: 5 studies compared PANSS-6 change scores with PANSS-30 change scores:

Studies 1–4/5 (Kolbaek et al., 2021b; Li et al., 2021; Ostergaard et al., 2018b, 2017) led to a pooled correlation of 0.76. *Hypothesis 8* was confirmed, and the 4 studies were hence rated "sufficient".

Study 5/5 (Lin et al., 2018) used Pearson x^2 could not detect any difference in change scores (p: 0.56–0.92). This finding confirmed our *hypothesis 7* and hence the study was rated "sufficient".

For *responsiveness,* 18 studies were rated as "sufficient", 2 studies as "insufficient" and 1 study as "indeterminate". Therefore, the overall *responsiveness* rating was "sufficient".

3.5. Grading the quality of evidence

For PANSS-6, the quality of evidence regarding structural validity, criterion validity and responsiveness was graded as "high", regarding measurement invariance, measurement error and construct validity (of the PANSS-6 total) as "moderate", regarding internal consistency and reliability as "low" and regarding content validity as "very low". The quality of evidence regarding the construct validity of the PANSS-6 subscales was found to be "moderate".

Detailed results on the quality of evidence together with the reasons for downgrading are presented in Table 3.

3.6. Recommendations

At the time of this review, there is neither enough evidence for "sufficient" *content validity* and *internal consistency* to categorize the PANSS-6 as (A) "can be recommended for use" nor is there high-quality evidence for an "insufficient" measurement property, which would lead to a categorization as (C) "should not be recommended for use".

Following COSMIN (Mokkink et al., 2018; Prinsen et al., 2018; Terwee et al., 2018), the PANSS-6 is categorized as (B), having potential to be recommended for use but requiring further research.

3.7. Feasibility

The greatest strength of the PANSS-6 is its short administration time of only 15–20 min (Kolbaek et al., 2018). The availability of the Simplified-Negative-and-Positive-Symptoms-Interview (SNAPSI) as a clinical interview guide is a further strength. The SNAPSI is available in 20 different languages (WCG Clinical Services, 2025). Furthermore, the PANSS-30 has been translated to over 40 different languages (Geck and Roithmeier et al., 2024), meaning that, as for PANSS-30, clear definitions of all items and the respective item anchors are available in many further languages as well (Kay, 2006). Kolbaek et al. and Nielsen et al. found that PANSS-6 ratings obtained with the SNAPSI do match the respective PANSS-6 ratings obtained with the SCI-PANSS (the structured interview designed for PANSS-30 assessments) legitimizing SNAPSIs use (Kolbaek et al., 2021a; Nielsen et al., 2022). The simple calculation of a total score is a further strength. However, users must not forget to

Table 3
COSMIN summary of findings.

Content validity	Summary or pooled result	Overall rating	Quality of evidence
PANSS-6	Relevance rating: +, Comprehensiveness rating: ?, Comprehensibility rating: + (ratings resulting from the reviewer rating)	Sufficient	Very Low (neither ClinROM development nor content validity study available)
Structural validity	Summary or pooled result	Overall rating	Quality of evidence
PANSS-6	CFI=0.994 for 2 factor model (total sample size: $n = 216$); IRT/ Rasch models for scalability accepted in 15/17 analyses (total sample size: $n = 2421$)	Sufficient	High
Internal consistency	Summary or pooled result	Overall rating	Quality of evidence
PANSS-6	Pooled Cronbach's α=0.695 (total sample size: $n = 644$)	Indeterminate	Low (multiple inadequate studies available)
Measurement invariance	Summary or pooled result	Overall rating	Quality of evidence
PANSS-6	4/5 analyses do not find Differential Item Functioning (DIF) for the group variables: age, gender, treatment, time (total sample size: $n = 2205$)	Sufficient	Moderate (multiple doubtful studies available)
Reliability	Summary or pooled result	Overall rating	Quality of evidence
PANSS-6	Pooled Interrater-reliability: 0.767[a] (total sample size: $n = 18$)	Sufficient	Low (multiple adequate studies available, total $n < 50$)
Measurement error	Summary or pooled result	Overall rating	Quality of evidence
PANSS-6	Standard Error of Measurement (SEM): 0.521 (total sample size: $n = 1073$)	Indeterminate	Moderate (only one study of adequate quality available)
Criterion validity	Summary or pooled result	Overall rating	Quality of evidence
PANSS-6	Pooled correlation coefficient: 0.808[b] (total sample size: n =2277)	Sufficient	High
Hypothesis testing	Summary or pooled result	Overall rating	Quality of evidence
PANSS-6	2 out of 2 hypotheses confirmed (total sample size $n = 250$)	Sufficient	Moderate (multiple studies of doubtful quality available)
PANSS-6 positive subscale	2 out of 2 hypotheses confirmed (total sample size $n = 250$)	Sufficient	Moderate (multiple studies of doubtful quality available)
PANSS-6 negative subscale	2 out of 2 hypotheses confirmed (total sample size $n = 250$)	Sufficient	Moderate (multiple studies of doubtful quality available)
Responsiveness	Summary or pooled result	Overall rating	Quality of evidence
PANSS-6	19 out of 21 hypotheses confirmed (total sample size $n = 9701$)	Sufficient	High

[a] pooled ICC and Gwet's agreement coefficents; [b] pooled Pearson and Spearman correlation coefficients.

deduct 6 points from the PANSS-6 total score when calculating percentage improvement to not underestimate response (Geck and Roithmeier et al., 2024; Leucht et al., 2007, 2009; Obermeier et al., 2010).

Weaknesses are that the PANSS-6 is, at least to date, only sold via licensing for research purposes (price upon request) (Pearson Assessments, 2023) and that potentially laborious rater trainings are required to improve the accuracy of ratings (Kolbaek et al., 2021a). Furthermore, the European Medicines Agency (EMA) requires the usage of the PANSS-30 for drug approval (European Medicines Agency, 2012). See Appendix 6.

4. Discussion

We present the first systematic review on the measurement properties of the PANSS-6. Based on 13 evaluation studies reviewed with the COSMIN methodology, the PANSS-6 is a good scale in terms of structural validity and responsiveness. Interrater reliability, criterion validity and construct validity were also rated sufficient, but the amount of evidence was relatively small. Evaluation studies are basically missing for content validity, cross-cultural validity and test-retest-reliability. The major advantage of the PANSS-6 compared to the full PANSS-30 is the shorter time needed for an assessment (15–20 min versus up to 60 min).

Structural validity was good in terms of scalability which is a major benefit over the PANSS-30 of which many items are not scalable. Bech has pointed out that to obtain valid measures of illness severity, scalability is essential (Bech, 2012; Lindenmayer, 2017). What is currently lacking for proof of good structural validity is an assessment of unidimensionality of the two subscales. It is likely because a confirmatory factor analysis (CFA) showed good fit for a model with two subscales and the items of these subscales loaded relatively reliably on the same factors of the PANSS-30 (Roithmeier et al., 2024), but it still needs to be formally tested.

In terms of internal consistency on average appropriate Cronbach's alpha values were achieved for the PANSS-6 *total score*. However, according to COSMIN unidimensionality is a prerequisite for being able to adequately assess internal consistency with Cronbach's alpha, but the PANSS-6 has been shown to be a two-dimensional scale. Internal consistency of the two subscales has not been tested yet.

We judged responsiveness to be sufficient according to COSMIN. Nevertheless, an important question is to what extent a short version such as PANSS-6 sufficiently covers schizophrenia symptoms. In fact, it covers four of the five symptoms defined as "core" according to DSM-5 (delusions, hallucinations, disorganized thinking, negative symptoms; except 'grossly disorganized or catatonic behavior')(Ostergaard et al., 2016). Of the four symptoms considered "core" by ICD-11 - delusions, hallucinations, thought disorder, and experiences of influence, passivity, or control – PANSS-6 does not cover the last (Keeley and Gaebel, 2018). Moreover, schizophrenia comprises more symptoms, in particular depressive or manic symptoms and cognitive deficits, which are mentioned in the DSM-5 and ICD-11 accompanying texts. The PANSS-6 covers only 3 (blunted affect, alogia and to some extent asocialty) of 5 domains suggested by the NIMH-MATRICS Consensus Statement on Negative symptoms (Marder and Kirkpatrick, 2014) but not avolition and anhedonia. In fact, the PANSS-30 also does not cover the latter two. Most factor analyses of the PANSS-30 suggest a five-factor solution (positive, negative, anxiety/depression, hostility/excitement, cognitive) (Roithmeier et al., 2024; Wallwork et al., 2012). (Takeuchi et al., 2023) have developed another short scale to cover these five factors and which might be a broader way to go. The PANSS-6 is not a diagnostic instrument but serves to identify symptom severity and its change. Nevertheless, we need to be aware that it does not cover all domains of the disorder it has been designed for limiting its usefulness for clinical practice. Ideally, a new short scale should be developed using appropriate methods (e.g. start out with a large question pool, involve experts and patients etc. (Mokkink et al., 2018; Prinsen et al., 2018; Terwee et al., 2018).

The first type of limitations are those of the status of evaluation of the PANSS-6. As it is a relatively new development (first publication in 2016) it can of course not be as extensively evaluated as the PANSS-30. Nevertheless, better assessment of interrater reliability, criterion validity and construct validity for which only some evidence is available is needed, and content validity, cross-cultural validity and test-retest-reliability have not been evaluated at all.

The second type of limitations are those related to our review process.

First, while the COSMIN framework (Mokkink et al., 2018; Prinsen et al., 2018; Terwee et al., 2018) offers a structured approach to evaluate the measurement properties of instruments such as the PANSS, it does not accommodate certain methodologies. For example, methods to assess measurement error such as standard error of measurement are not considered by COSMIN, therefore, such studies yielded an indeterminate rating.

In a similar vein some articles (Hieronymus et al., 2023, 2021; Lin et al., 2018; Pagsberg et al., 2022) that assessed responsiveness, additionally investigated how *initial symptom severity and early changes* in PANSS-6 scores relate to later treatment outcomes, including response, remission, and dropout rates. We excluded these analyses (not the whole articles) because they combined validation of the scale with validation of the early prediction of response concept. Nevertheless, as responsiveness was judged sufficient anyhow, inclusion would not have changed the COSMIN judgement.

Third, our choice of the PANSS-30 as the gold standard for the psychometric evaluation of the PANSS-6 is questionable because, overall, its psychometric properties are not ideal (Geck and Roithmeier et al., 2024). However, for the measurement properties where it was used as a comparator in this review –criterion validity and responsiveness – the PANSS-30 is sufficiently good and not considering the PANSS-30 as gold standard for criterion validity would not have changed the overall results of our review.

Fourth, although COSMIN is a highly elaborated and operationalized framework to systematically review the measurement properties of rating scales, there is a certain degree of subjectivity.

Fifth, many PANSS-6 ratings in our included publications were extracted from PANSS-30 assessments rather than being collected individually. Thus, the assessments were not independent. More studies in which the PANSS-6 and the PANSS-30 are assessed independently would be desirable.

5. Conclusion

According to COSMIN the PANSS-6 is a scale "with potential to be recommended for use". If internal consistency were further evaluated and found sufficient, the recommendation could become "recommended for use". The main advantage of the PANSS-6 over the PANSS-30 is the much shorter application time with no important loss in responsiveness making it useful for clinical trials. The PANSS-6 is elegant in the way that it is based on the widely accepted remission criteria by (Andreasen et al., 2005) which are thus covered simultaneously in a PANSS-6 assessment. Fewer symptoms of schizophrenia are covered than by the PANSS-30. Nevertheless, even the PANSS-30 is not ideal in this regard, because it e.g. does not cover all relevant negative symptom domain or cognition. Therefore, there is also room for new, comprehensive and short schizophrenia rating scales.

Author contributions

The authors confirm contribution to the paper as follows: study conception and design: S.G., M.R., S.L.; data collection: S.G. (first rater), M.R. (second rater); analysis and interpretation of results: S.G. (first rater), M.R. (second rater), S.L. (third rater); statistical consultation: M. B.; draft manuscript preparation: S.G., M.R.; revision for important intellectual content: S.L., M.B., S.W., L.W., J.P., J.D.; The work will be part

of the doctoral thesis of S.G.; All authors reviewed the results and approved the final version of the manuscript. All authors have agreed to be personally accountable for the author's own contributions and to ensure that questions related to the accuracy or integrity of any part of the work, are appropriately investigated, resolved, and the resolution documented in the literature.

Role of the funding source

There was no specific funding source for this research.

Declaration of competing interest

In the last 3 years, Dr. Leucht has received honoraria for advising, consulting, and for lectures and/or for educational material from Angelini, Apsen, Boehringer Ingelheim, Eisai, Ekademia, Gedeon-Richter, Janssen, Karuna, Kynexis, Lundbeck, Medichem, Medscape, Mitsubishi, Neurotorium, Otsuka, Novo Nordisk, Recordati, Rovi, and Teva. All other authors declare that they have no conflicts of interest relevant to this research.

Acknowledgement

We are grateful to Lidwine B. Mokkink from COSMIN for support in resolving ambiguities regarding the applied methodology.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT-4 in order to improve readability. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.euroneuro.2025.02.008.

References

Andreasen, N.C., Carpenter Jr., W.T., Kane, J.M., Lasser, R.A., Marder, S.R., Weinberger, D.R., 2005. Remission in schizophrenia: proposed criteria and rationale for consensus. Am. J. Psychiatry 162, 441–449.

Association, A.P., 2013. Diagnostic and statistical manual of mental disorders (5th ed.).

Baandrup, L., Allerup, P., Nielsen, M.O., During, S.W., Bojesen, K.B., Leucht, S., Galderisi, S., Mucci, A., Bucci, P., Arango, C., Diaz-Caneja, C.M., Dazzan, P., McGuire, P., Demjaha, A., Ebdrup, B.H., Fleischhacker, W.W., Kahn, R.S., Glenthoj, B.Y., 2022. Scalability of the positive and negative syndrome scale in first-episode schizophrenia assessed by Rasch models. Acta Psychiatr. Scand. 146, 21–35.

Bech, P., 2012. Clinical Psychometrics. Wiley Blackwell, Oxford.

Borenstein, M., Hedges, L., Higgins, J., Rothstein, H., 2009. An introduction to meta-analysis. Introduct. Meta-Anal. 19.

European Medicines Agency, 2012. Guideline on clinical investigation of medicinal products, including depot preparations, in the treatment of schizophrenia - revision 1.

Geck, S.A., & Roithmeier, M.A.; Bühner, M., Wehr, S.; Weigel, L., Priller, J., Leucht, S., 2024. COSMIN systematic review and meta-analysis of the measurement properties of the Positive and negative Syndrome scale (PANSS) [manuscript under review].

GRADE, 2013. GRADE Handbook - Handbook For Grading the Quality of Evidence and the Strength of Recommendations Using the GRADE Approach.

Hieronymus, F., Correll, C.U., Østergaard, S.D., 2023. Initial severity of the Positive and Negative Syndrome Scale (PANSS)-30, its main subscales plus the PANSS-6, and the relationship to subsequent improvement and trial dropout: a pooled participant-level analysis of 18 placebo-controlled risperidone and paliperidone trials. Transl. Psychiatry 13, 191.

Hieronymus, F., Kolbaek, P., Correll, C.U., Ostergaard, S.D., 2021. Antipsychotic-placebo separation on the PANSS-6 subscale as compared to the PANSS-30: a pooled participant-level analysis. NPJ. Schizophr. 7, 41.

Huhn, M., Nikolakopoulou, A., Schneider-Thoma, J., Krause, M., Samara, M., Peter, N., Arndt, T., Bäckers, L., Rothe, P., Cipriani, A., Davis, J., Salanti, G., Leucht, S., 2019. Comparative efficacy and tolerability of 32 oral antipsychotics for the acute treatment of adults with multi-episode schizophrenia: a systematic review and network meta-analysis. Lancet 394, 939–951.

Kay, S.R., Fiszbein, A., Opler, L.A., 1987. The positive and negative syndrome scale (PANSS) for schizophrenia. Schizophr. Bull. 13, 261–276.

Kay, S.R., Opler, L.A., Fiszbein, A., 2006. Positive and Negative Syndrome Scale (PANSS): Manual. Multi-Health Systems, Inc.

Keeley, J.W., Gaebel, W., 2018. Symptom rating scales for schizophrenia and other primary psychotic disorders in ICD-11. Epidemiol. Psychiatr. Sci. 27, 219–224.

Kolbaek, P., Blicher, A.B., Buus, C.W., Feller, S.G., Holm, T., Dines, D., O'Leary, K.M., Sorensen, R.S., Opler, M., Correll, C.U., Mors, O., Bech, P., Ostergaard, S.D., 2018. Inter-rater reliability of ratings on the six-item Positive and Negative Syndrome Scale (PANSS-6) obtained using the Simplified negative and Positive Symptoms Interview (SNAPSI). Nord. J. Psychiatry 72, 431–436.

Kolbaek, P., Dines, D., Hansen, J., Opler, M., Correll, C.U., Mors, O., Ostergaard, S.D., 2021a. Standardized training in the rating of the six-item Positive and Negative Syndrome scale (PANSS-6). Schizophr. Res. 228, 438–446.

Kolbaek, P., Dines, D., Holm, T., Blicher, A.B., Sorensen, R.D., O'Leary, K.M., Feller, S.G., Buus, C.W., Nielsen, C.M., Opler, M., Mors, O., Correll, C.U., Ostergaard, S.D., 2021b. Clinical validation of ratings on the six-item positive and negative syndrome scale obtained via the simplified negative and positive symptoms interview. J. Psychopharmacol. 35, 1081–1090.

Leucht, S., Davis, J.M., Engel, R.R., Kane, J.M., Wagenpfeil, S., 2007. Defining 'response' in antipsychotic drug trials: recommendations for the use of scale-derived cutoffs. Neuropsychopharmacology 32, 1903–1910.

Leucht, S., Davis, J.M., Engel, R.R., Kissling, W., Kane, J.M., 2009. Definitions of response and remission in schizophrenia: recommendations for their use and their presentation. Acta Psychiatr. Scand. Suppl. 7–14.

Li, L., Ma, H., Wang, X., Meng, E., 2021. Validation of Chinese version of positive and negative syndrome scale-6 in clinical setting: a preliminary study. Psychiatry Clin. Psychopharmacol. 31, 386–391.

Lin, C.H., Lin, H.S., Lin, S.C., Kuo, C.C., Wang, F.C., Huang, Y.H., 2018. Early improvement in PANSS-30, PANSS-8, and PANSS-6 scores predicts ultimate response and remission during acute treatment of schizophrenia. Acta Psychiatr. Scand. 137, 98–108.

Lindenmayer, J.P., 2017. Are shorter versions of the positive and negative syndrome scale (PANSS) doable? A critical review. Innov. Clin. Neurosci. 14, 73–76.

Marder, S.R., Kirkpatrick, B., 2014. Defining and measuring negative symptoms of schizophrenia in clinical trials. Eur. Neuropsychopharmacol. 24, 737–743.

Mokkink, L.B., de Vet, H.C.W., Prinsen, C.A.C., Patrick, D.L., Alonso, J., Bouter, L.M., Terwee, C.B., 2018. COSMIN risk of bias checklist for systematic reviews of patient-reported outcome measures. Qual. Life Res. 27, 1171–1179.

Nielsen, C.M., Kolbaek, P., Dines, D., Pedersen, M.L., Danielsen, A.A., Holmgard, C., Wissing, S., Esbol, A.M., Fuglsang, N.F.B., Nguyen, T.D., Mors, O., Opler, M., Correll, C.U., Ostergaard, S.D., 2022. Validation of ratings on the six-item positive and negative syndrome scale obtained via the simplified negative and positive symptoms Interview among outpatients with schizophrenia. J. Psychopharmacol. 36, 1208–1217.

Obermeier, M., Mayr, A., Schennach-Wolff, R., Seemüller, F., Möller, H.J., Riedel, M., 2010. Should the PANSS be rescaled? Schizophr. Bull. 36, 455–460.

Ostergaard, S.D., Foldager, L., Mors, O., Bech, P., Correll, C.U., 2018a. The validity and sensitivity of PANSS-6 in the clinical antipsychotic trials of intervention effectiveness (CATIE) study. Schizophr. Bull. 44, 453–462.

Ostergaard, S.D., Foldager, L., Mors, O., Bech, P., Correll, C.U., 2018b. The validity and sensitivity of PANSS-6 in treatment-resistant schizophrenia. Acta Psychiatr. Scand. 138, 420–431.

Ostergaard, S.D., Lemming, O.M., Mors, O., Correll, C.U., Bech, P., 2016. PANSS-6: a brief rating scale for the measurement of severity in schizophrenia. Acta Psychiatr. Scand. 133, 436–444.

Ostergaard, S.D., Opler, M.G.A., Correll, C.U., 2017. Bridging the measurement gap between research and clinical care in schizophrenia: positive and negative syndrome scale-6 (PANSS-6) and othassessments based on the simplified negative and positive symptoms interview (SNAPSI). Innov. Clin. Neurosci. 14, 11–12.

Pagsberg, A.K., Krogmann, A., Jeppesen, P., von Hardenberg, L., Klauber, D.G., Jensen, K.G., Rudå, D., Decara, M.S., Jepsen, J.R.M., Fagerlund, B., Fink-Jensen, A., Correll, C.U., Galling, B., 2022. Early antipsychotic nonresponse as a predictor of nonresponse and nonremission in adolescents with psychosis treated with Aripiprazole or Quetiapine: results from the TEA trial. J. Am. Acad. Child Adolesc. Psychiatry 61, 997–1009.

Pearson Assessments, 2023. Positive and negative syndrome scale. https://www.pearsonassessments.com/store/usassessments/en/Store/Professional-Assessments/Personality-%26-Biopsychosocial/Positive-and-Negative-Syndrome-Scale/p/P100025000.html (accessed 23.08.2024).

Prinsen, C.A.C., Mokkink, L.B., Bouter, L.M., Alonso, J., Patrick, D.L., de Vet, H.C.W., Terwee, C.B., 2018. COSMIN guideline for systematic reviews of patient-reported outcome measures. Qual. Life Res. 27, 1147–1157.

Roithmeier M.A., Geck S.A., Peter N., Leucht S., 2024a. Systematic review of the Positive and Negative Syndrome Scale following the COSMIN standard.

Roithmeier, M.A., Geck, S.A., Bühner, M., Wehr, S., Weigel, L., Priller, J., Davis, J.M., Leucht, S., 2024b. COSMIN review of the PANSS Marder factor solution and other factor models in people with schizophrenia [manuscript under review].

Santor, D.A., Ascher-Svanum, H., Lindenmayer, J.-P., Obenchain, R.L., 2007. Item response analysis of the positive and Negative Syndrome scale. BMC. Psychiatry 7, 66.

WCG Clinical Services, 2025. Download SNAPSI resources. https://www.wcgclinical.com/insights/download-snapsi-resources/(accessed 28.01.2025).

Singh, M.M., Kay, S.R., 1975. A comparative study of haloperidol and chlorpromazine in terms of clinical effects and therapeutic reversal with benztropine in schizophrenia. Theoretical implications for potency differences among neuroleptics. Psychopharmacologia 43, 103–113.

Takeuchi, H., Lee, J., Fervaha, G., Agid, O., Remington, G., 2023. Second version of brief evaluation of psychosis symptom domains (BE-PSD-V2.0). Schizophr. Res. 254, 176–177.

Terwee, C.B., Prinsen, C.A.C., Chiarotto, A., Westerman, M.J., Patrick, D.L., Alonso, J., Bouter, L.M., de Vet, H.C.W., Mokkink, L.B., 2018. COSMIN methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study. Qual. Life Res. 27, 1159–1170.

Wallwork, R.S., Fortgang, R., Hashimoto, R., Weinberger, D.R., Dickinson, D., 2012. Searching for a consensus five-factor model of the Positive and Negative Syndrome Scale for schizophrenia. Schizophr. Res. 137, 246–250.

Weigel, L., Wehr, S., Galderisi, S., Mucci, A., Davis, J., Giordano, G.M., Leucht, S., 2023. The Brief negative Symptom Scale (BNSS): a systematic review of measurement properties. Schizophrenia (Heidelb) 9, 45.