TECHNICAL COMMENTARY

SCHIZOPHRENIA LIBRARY

Outcome assessment tools

Introduction

Standardised assessment tools are vital for measuring a range of variables including symptoms, functioning and quality of life. They are often used within a controlled research environment, but high-quality assessment tools are also useful in practice for both clinical management and outcome prediction.

The quality of assessment tools can be measured in various ways. ‘Reliability’ refers to the reproducibility of an instrument’s results across different assessors, settings, and times. ‘Construct validity’ is the extent to which an instrument measures the theoretical construct it was designed to measure. This involves ‘convergent validity’, which is the degree of correlation between different scales measuring the same construct, confirming they are measuring the same thing; and ‘divergent validity’, which is the lack of correlation between scales measuring different constructs, confirming that they are measuring different things. Similarly, ‘known groups’ validity’ is the extent to which an instrument can demonstrate different scores for groups known to vary on the variables being measured. ‘Content validity’ is the extent to which each individual item on a scale represents the construct being measured, and ‘internal consistency’ is the degree of correlation between individual items within a scale.

‘Predictive validity’ refers to sensitivity, which is the proportion of correctly identified positives, and specificity, which is the proportion of correctly identified negatives. Sensitivity and specificity are measured by comparing an instrument’s results with known ‘gold standard’ results. ‘Responsiveness’ is the extent to which an instrument can detect clinically significant or practically important changes over time, and ‘area under the curve’ (AUC) is a global measure of test performance.

Method

We have included only systematic reviews with detailed literature search, methodology, and inclusion/exclusion criteria that were published in full text, in English, from the year 2000. Reviews were identified by searching the databases MEDLINE, EMBASE, and PsycINFO. Reviews with pooled data are prioritized for inclusion. Reviews reporting fewer than 50% of items on the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA[1]) checklist have been excluded from the library. The evidence was graded guided by the Grading of Recommendations Assessment, Development and Evaluation (GRADE) Working Group approach[2]. The resulting table represents an objective summary of the available evidence, although the conclusions are solely the opinion of staff of NeuRA (Neuroscience Research Australia).

Results

We found 18 systematic reviews that met our inclusion criteria[3-20].

- Moderate to high quality evidence finds good predictive value of < 20% reduction on the Brief Psychiatric Rating Scale or the Positive and Negative Syndrome Scale after 2 weeks of antipsychotic treatment for predicting non-response to treatment at 4-12 weeks. Better specificity was associated with shorter trial duration, higher baseline illness severity, and shorter illness duration.
- Moderate quality evidence finds the Brief Psychiatric Rating Scale can be factored into five discrete components, comprising positive, negative, and affective symptoms, resistance (hostility) and activation (excitement). The Positive and Negative Syndrome Scale showed a similar factor structure to the Brief Psychiatric Rating Scale but included a larger number of items in the negative symptom factor and enough items for a discrete disorganisation factor.

- Moderate to high quality evidence suggests good predictive validity of the Historical, Clinical and Risk Management-20 scale for predicting aggression in psychiatric facilities. The best predictive efficacy was for samples containing higher proportions of people with schizophrenia, women, and Caucasians.
- Moderate to low quality evidence suggests the McNiel-Binder Violence Screening Checklist, and the Brøset Violence Checklist may also be effective for predicting aggression or violence. The Violence Risk Appraisal Guide had poor predictive validity in people with schizophrenia living in the community.
- Moderate quality evidence suggests good inter-rater reliability and small predictive validity for tools assessing duration of untreated psychosis, psychosis onset and treatment onset.
- Moderate to low quality evidence suggests the Recovery Assessment Scale has the best psychometric properties for measuring personal recovery in schizophrenia. It is rated as having good construct validity, content validity, internal consistency, test-retest reliability, administrator-friendliness, and has been translated to languages other than English. However, its user-friendly rating is poor. Other scales rating personal recovery with reasonable psychometric properties include the Self-Identified Stage of Recovery scale, which has good construct and content validity, good internal consistency (but poor test-retest reliability), good user-friendliness, and has been translated to languages other than English. The Mental Health Recovery Measure has good content validity (but poor construct validity), and good internal consistency and test-re-test validity.
- Moderate to low quality evidence suggests reliability is good for instruments assessing comorbid depressive symptoms in people with schizophrenia; Brief Psychiatric Rating Scale-Depression, Positive and Negative Syndrome Scale-Depression, Hamilton Rating Scale for Depression, Montgomery Asberg Depression Rating Scale, Calgary Depression Scale for Schizophrenia, and Beck Depression Inventory. The Montgomery Asberg Depression Rating Scale showed a medium-sized correlation with negative symptoms of schizophrenia, and the Hamilton Rating Scale for Depression showed a medium-sized correlation with extrapyramidal symptoms (measured using various scales), suggesting poor divergent validity for these instruments. The best concurrent validity indices were reported for the Calgary Depression Scale for Schizophrenia, and the Montgomery Asberg Depression Rating Scale. The highest ranges for sensitivity and specificity were reported for the Calgary Depression Scale for Schizophrenia.
- For anxiety symptoms, moderate quality evidence suggests the Beck Anxiety Index, Depression Anxiety Stress Scale or Scale of Anxiety Evaluation in Schizophrenia for general screening, and the DSM-based Generalised Anxiety Disorder Symptoms Severity Scale, Liebowitz Social Anxiety Scale, Obsessive-Compulsive Inventory, Psychological Stress Index, Perseverative Thinking Questionnaire, and Yale-Brown Obsessive Compulsive Scale for anxiety symptoms.
- Moderate to low quality evidence suggests good ‘known groups’ validity for the Short Form health survey-36, but inconsistent convergent validity and poor responsiveness.
- Moderate to low quality evidence suggests 62% of studies reviewed had incorrectly calculated ratios using the Positive and Negative Syndrome Scale, potentially resulting in inadvertently lower response rates.
- Moderate to low quality evidence suggests the use of a modified Scale to Assess Unawareness of Mental Disorder may

compromise the psychometric properties of the scale, lead to erroneous conclusions, and prevent comparison of results across studies.

- Moderate to high quality evidence suggests small relationships between self-report and clinician-rated, performance-based and clinician rated, and amotivation self-report and amotivation clinician-rated assessments of motivation.
- Moderate quality evidence finds a medium-sized effect of increased detection of symptomology in assessments conducted in the mother language rather than the acquired language.
- Moderate to low quality evidence suggests good internal consistencies for the Visual Jokes task, Faux Pas task, Reading the Mind in the Eyes Test, and the Moving Shapes task. The Hinting task and False Belief picture Sequencing showed moderate internal consistencies. Good test-retest reliabilities were reported for the Hinting task and the Faux Pas task. The Story test had moderate test-retest reliability. The False Belief stories task and the second-order False Belief stories task had poor reliability. The Reading the Mind in the Eyes Test had inconsistent reliability measures.

Bakkour N, Samp J, Akhras K, El Hammi E, Soussi I, Zahra F, Duru G, Kooli A, Toumi M

Systematic review of appropriate cognitive assessment instruments used in clinical trials of schizophrenia, major depressive disorder and bipolar disorder

Psychiatry Research 2014; 216: 291-302

View review abstract online

Comparison	**Identification of appropriate scales to measure cognition in people with schizophrenia according to the Measurement and Treatment Research to Improve Cognition in Schizophrenia (MATRICS) initiative.** **This initiative identified five criteria for scale appropriateness:** **1. Test–retest reliability** **2. Utility as a repeated measure** **3. Relationship to functional outcomes** **4. Potential changeability in response to pharmacological agents** **5. Tolerability and practicality for a clinical setting**
Summary of evidence	**Moderate to low quality evidence (direct, unable to assess consistency or precision) suggests a range of cognitive scales are appropriate for measuring cognition in people with schizophrenia.**

Appropriate scales to measure cognition

The following appropriate measurement scales were identified;

Brief Assessment of Cognition in Schizophrenia

Beck Cognitive Insight Scale

Brief Visualspatial Memory Test Revised

Cambridge Neuropsychological Test Automated Battery

Coping Attitude Scale

Category Fluency: Animal naming

Clinical Global Impression of Cognition in Schizophrenia

CogState Schizophrenia Battery Computerized neurocognitive battery Continuous Performance Test-Identical Pairs California Verbal Learning Test Hypomanic Attitudes and Positive Predictions Inventory Hopkins Verbal Learning Test-R scores IntegNeuro Letter Number Span Mindstreams Computerized Cognitive Test Battery MATRICS Consensus Cognitive Battery Mayer-Saovey-Caruso Emotional Intelligence Test Neuropsychological assessment battery Repeatable Battery for the Assessment of Neuropsychological Status Schizophrenia Communication Disorder Scale Schizophrenia Cognition Rating Scale Skills of Cognitive Therapy Trail Making Test A Wechsler Adult Intelligence Scale-Revised Wechsler Memory Scale Wechsler Intelligence Scale for Children Wide Range Achievement Test	
Consistency in results‡	Unable to assess; no measure of consistency is reported
Precision in results§	Unable to assess; no measure of precision is reported
Directness of results‖	Direct

Cavelti M, Kyrgic S, Beck EM, Kossowsky J, Vauth R

Assessing recovery from schizophrenia as an individual process. A review of self-report instruments

European Psychiatry 2012; 27: 19-32

View review abstract online

Comparison	**Psychometric properties of instruments assessing self-reported, clinical and functional outcomes of personal recovery from schizophrenia.**
Summary of evidence	**Moderate to low quality evidence (unclear sample sizes, direct, unable to assess consistency or precision) suggests that of the available instruments measuring personal recovery, the Recovery Assessment Scale (RAS) had the best psychometric properties, with good construct validity & content validity, internal consistency, test-retest reliability, administrator-friendliness, and was translated to languages other than English. However, its user-friendly rating was poor.** **Other scales with reasonable psychometric properties include; the Self-Identified Stage of Recovery scale (SISR), which had good construct and content validity, good internal consistency (but poor test-retest reliability), good user-friendliness, and was translated to languages other than English. The Mental Health Recovery Measure (MHRM) had good content validity (but poor construct validity), and good internal consistency and test-re-test validity.**
Validity	
Consumer Recovery Outcomes System (CROS 3.0): indeterminate construct, good content Illness Management and Recovery Scale (IMR): poor construct, good content Modified Engulfment Scale (MES): good construct, poor content Mental Health Recovery Measure (MHRM): poor construct, good content Ohio Outcomes System: poor construct, good content Patient Outcomes Research Team Scale (PORT): no studies assessing construct, poor content Psychosis Recovery Inventory (PRI): poor construct, good content Recovery Assessment Scale (RAS): good construct, good content Recovery Attitudes Questionnaire - 7 (RAQ-7): no studies assessing construct, good content Recovery Process Inventory (RPI): poor construct, good content Recovery Style Questionnaire (RSQ): indeterminate construct, good content Stage of Recovery Instrument (STORI): indeterminate construct, good content Self-Identified Stage of Recovery (SISR): good construct, good content	
Reliability	

CROS 3.0: good internal consistency, indeterminate test-retest reliability IMR: poor internal consistency, good test-retest reliability MES: indeterminate internal consistency, no studies assessing test-retest reliability MHRM: good internal consistency, good test-retest reliability Ohio Outcomes: indeterminate internal consistency, no studies assessing test-retest reliability PORT: good internal consistency, no studies assessing test-retest reliability PRI: indeterminate internal consistency, poor test-retest reliability RAS: good internal consistency, good test-retest reliability RAQ-7: good internal consistency, poor test-retest reliability RPI: poor internal consistency, poor test-retest reliability RSQ: indeterminate internal consistency, good test-retest reliability STORI: indeterminate internal consistency, good test-retest reliability SISR: good internal consistency, poor test-retest reliability	
Issues of application	
CROS 3.0: indeterminate user and administrator friendliness, no translations IMR: indeterminate user and good administrator friendliness, good translations MES: indeterminate user and poor administrator friendliness, no translations MHRM: indeterminate user and administrator friendliness, no translations Ohio Outcomes: poor user and good administrator friendliness, indeterminate translations PORT: poor user and good administrator friendliness, no translations PRI: indeterminate user and administrator friendliness, indeterminate translations RAS: poor user and good administrator friendliness, good translations RAQ-7: good user and indeterminate administrator friendliness, no translations RPI: indeterminate user and administrator friendliness, no translations RSQ: indeterminate user and good administrator friendliness, no translations STORI: poor user and good administrator friendliness, indeterminate translations SISR: good user and indeterminate administrator friendliness, good translations	
Consistency in results	Unable to assess; no measure of consistency is reported
Precision in results	Unable to assess; no measure of precision is reported
Directness of results	Direct

Dumas R, Baumstarck K, Michel P, Lançon C, Auquier P, Boyer L **Systematic Review Reveals Heterogeneity in the Use of the Scale to Assess Unawareness of Mental Disorder (SUMD)** **Current Psychiatry Reports 2013; 15: 361** View review abstract online	
Comparison	**Use of the Scale to Assess Unawareness of Mental Disorder.**
Summary of evidence	**Moderate to low quality evidence (unclear sample size, direct, unable to assess consistency or precision) finds the use of a modified SUMD may compromise the psychometric properties of the scale, lead to erroneous conclusions, and prevent comparison of results across studies.**
Use of the Scale to Assess Unawareness of Mental Disorder	
100 studies were included in the review. Authors report that the SUMD is one of the most widely used instruments to measure insight, and it has satisfactory psychometric properties. However, the SUMD was rarely used in its entirety and calculation of insight scores was highly variable. The use of a modified SUMD may compromise the psychometric properties of the scale, lead to erroneous conclusions, and prevent comparison of results across studies.	
Consistency in results	Unable to assess; no measure of consistency is reported
Precision in results	Unable to assess; no measure of precision is reported
Directness of results	Direct

Erkoreka L, Ozamiz-Etxebarria N, Ruiz O, Ballesteros J **Assessment of psychiatric symptomatology in bilingual psychotic patients: A systematic review and meta-analysis** **International Journal of Environmental Research and Public Health 2020; 17(11): 1-11** View review abstract online	
Comparison	**Assessments conducted in mother vs. acquired language.**

Summary of evidence	**Moderate quality evidence (small to medium sample size, inconsistent, precise, direct) finds a medium-sized effect of increased detection of symptomology in assessments conducted in the mother language rather than the acquired language.**
Assessments of symptoms	
A medium-sized effect of increased detection of symptomology in assessments conducted in the mother language rather than the acquired language; 4 studies, N = 283, SMD = 0.44, 95%CI 0.19 to 0.69, *p* = 0.0006, I^2 = 90%	
Consistency in results	Inconsistent
Precision in results	Precise
Directness of results	Direct

Lako LM, Bruggeman R, Knegtering H, Wiersma D, Schoevers RA, Slooff CJ, Taxis K

A systematic review of instruments to measure depressive symptoms in patients with schizophrenia

Journal of Affective Disorders 2012; 140: 38-47

View review abstract online

Comparison	**Psychometric properties of instruments that measure depressive symptoms in people with schizophrenia.**
Summary of evidence	**Six instruments were assessed: BPRS-D, PANSS-D, HAM-D, MADRS, CDSS, and BDI.** **Moderate to low quality evidence (unclear sample sizes, direct, some imprecision, unable to assess consistency) suggests reliability was good for all instruments. The highest ranges for sensitivity and specificity were reported for the CDSS. The MADRS and HAM-D showed poor divergent validity, with medium correlations with negative symptoms (various measures), and the HAM-D showed medium correlations with extrapyramidal symptoms (various measures). The best concurrent validity indices were reported for the CDSS and MADRS.**

Reliability **Measured by Cronbach's alpha**
Reliability was good for all instruments; Brief Psychiatric Rating Scale (BPRS-D): internal consistency 0.67, inter-rater 0.74, test-retest 0.72 (2 studies) Positive and Negative Syndrome Scale (PANSS-D): internal consistency 0.77, inter-rater 0.80 (2 studies) Hamilton Rating Scale for Depression (HAM-D): internal consistency 0.75, inter-rater 0.94, test-retest 0.75 (5 studies) Montgomery Asberg Depression Rating Scale (MADRS): internal consistency 0.91, inter-rater 0.81, test-retest 0.71 (3 studies) Calgary Depression Scale for Schizophrenia (CDSS): internal consistency 0.82, inter-rater 0.86, test-retest 0.83 (13 studies) Beck Depression Inventory (BDI): internal consistency 0.90 (2 studies)
Divergent validity **Negative and extrapyramidal symptoms (EPS) were measured by the Affective Flattening Scale, Scale for the Assessment of Negative Symptoms, negative subscale of the PANSS, negative subscale of the BPRS, Psychomotor Retardation Scale and Rating Scale for Extrapyramidal Side Effects**
The CDSS, BDI, PANSS-D and BPRS-D showed acceptably low divergent effects relative to either negative symptoms or EPS, indicating high specificity for measuring depression. The MADRS showed poor divergent validity, with a medium size correlation with both negative symptoms and EPS, and the HAM-D showed a medium size correlation with EPS; BPRS-D: negative symptoms *r* = 0.00, 95%CI −0.11 to 0.10; EPS *r* = 0.14 95%CI 0.07 to 0.21 PANSS-D: negative symptoms *r* = 0.19, 95%CI −0.11 to 0.41; EPS *r* = 0.07, 95%CI 0.01 to 0.20 HAM-D: negative symptoms *r* = 0.18, 95%CI 0.02 to 0.45; EPS *r* = 0.40, 95%CI 0.02 to 0.79 MADRS: negative symptoms *r* = 0.36, 95%CI 0.12 to 0.51; EPS *r* = 0.52, 95%CI 0.16 to 0.86 CDSS: negative symptoms *r* = 0.10, 95%CI −0.24 to 0.54; EPS *r* = 0.26, 95%CI 0.07 to 0.42 BDI: negative symptoms *r* = 0.10, 95%CI −0.11 to 0.21; EPS *r* = 0.23, 95%CI not reported
Concurrent validity **Measured by correlations with each other depression scale**
The highest concurrent validity indices were found for the CDSS and MADRS; BPRS-D: PANSS-D *r* = 0.23, HAMD *r* = 0.66, MADRS *r* = 0.66, CDSS *r* = 0.79, BDI *r* = 0.64 Pooled across all scales *r* = 0.60, 95%CI 0.17 to 0.87

PANSS-D: HAMD *r* = 0.62, MADRS *r* = 0.72, CDSS *r* = 0.66, BDI *r* = 0.49 Pooled across all scales *r* = 0.54, 95%CI 0.17 to 0.87 HAMD: MADRS *r* = 0.80 CDSS *r* = 0.74 BDI *r* = 0.57 Pooled across all scales *r* = 0.68, 95%CI 0.26 to 0.90 MADRS: CDSS *r* = 0.81 Pooled across all scales *r* = 0.75, 95%CI 0.56 to 0.90 CDSS: BDI *r* = 0.83 Pooled across all scales *r* = 0.77, 95%CI 0.26 to 0.90 BDI: Pooled across all scales *r* = 0.63, 95%CI 0.44 to 0.90	
Predictive validity **Measures as sensitivity and specificity for predicting a major depressive episode**	
The highest ranges for sensitivity and specificity were found for the CDSS; PANSS-D: sensitivity 78%, 95%CI 74% to 81%, specificity 85% 95%CI 79% to 90%, cut-off value ≥5; ≥10 HAMD: sensitivity 79% 95%CI 67% to 91% specificity 83% 95%CI 81% to 84% cut-off value ≥12 MADRS: sensitivity 81%, specificity 81%, Cut-off value ≥11 CDSS: sensitivity 88% 95%CI 67% to 100%, specificity 88% 95%CI 74% to 97%, cut-off value ≥5; ≥6; ≥9 BDI: sensitivity 72%, specificity 77%, cut-off value ≥25	
Consistency in results	Unable to assess; no measure of consistency is reported
Precision in results	Divergent validity is mostly precise, concurrent validity is mostly imprecise. Predictive validity appears precise. Unable to assess reliability
Directness of results	Direct

Luther L, Firmin RL, Lysaker PH, Minor KS, Salyers MP

A meta-analytic review of self-reported, clinician-rated, and performance-based motivation measures in schizophrenia: Are we measuring the same "stuff"?

Clinical Psychology Review 2018; 61: 24-37

View review abstract online	
Comparison	**Relationship between different methods of motivation assessment in people with schizophrenia.**
Summary of evidence	**Moderate to high quality evidence (medium to large samples, some inconsistency, precise, direct) suggests small relationships between self-report and clinician-rated, performance-based and clinician rated, and amotivation self-report and amotivation clinician-rated assessments of motivation.**
Motivation **Intrinsic Motivation Inventory for Schizophrenia Research** **Amotivation subscale of the Positive and Negative Syndrome Scale** **Effort Expenditure for Rewards Task**	
Significant, small relationships between; Self-report and clinician-rated 33 studies, N = 2270, *r* = 0.27, 95%CI 0.19 to 0.35, *p* < 0.001, I^2 = 73%, *p* < 0.001 Performance-based and clinician-rated 11 studies, N = 445, *r* = 0.21, 95%CI 0.10 to 0.32, *p* < 0.001, I^2 = 13%, *p* > 0.05 Amotivation self-report and amotivation clinician-rated 23 studies, N = 1847, *r* = 0.34, 95%CI 0.24 to 0.43, *p* < 0.001, I^2 = 77%, *p* < 0.001 *There were no relationships between;* Self-report and performance-based 2 studies, N = 128, *r* = -0.001, 95%CI -0.21 to 0.21, *p* > 0.05, I^2 = 21%, *p* > 0.05 Intrinsic motivation self-report and intrinsic motivation clinician-rated 4 studies, N = 209, *r* = 0.16, 95%CI -0.12 to 0.42, *p* > 0.05, I^2 = 75%, *p* < 0.01	
Consistency in results	Inconsistent for self-report/clinician-rated, amotivation self-report/ amotivation clinician-rated, and intrinsic motivation self-report/intrinsic motivation clinician-rated
Precision in results	Precise
Directness of results	Direct

Obermeier M, Schennach-Wolff R, Meyer S, Möller H, Riedel M, Krause D, *Seemüller F*

Is the PANSS used correctly? A systematic review

BMC Psychiatry 2011; (11): 113

View review abstract online

Comparison	**Assessment of the use of the PANSS instrument in scientific research articles.**
Summary of evidence	**Moderate to low quality evidence (unclear sample size, direct, unable to assess consistency or precision) suggests 62% of studies reviewed had incorrectly calculated ratios using the PANSS, potentially resulting in inadvertently lower response rates.**
Correct calculation of proportions	
The PANSS is a 30-item interval scale, with the possible score for individual items ranging from 1-7. This implies that computations of ratios (e.g., % change from baseline) are not appropriate without first subtracting the minimum (e.g., 30 for the total score), to give a score range starting from zero. 24/39 (62%) of studies reviewed had used incorrect calculations using the PANSS, potentially resulting in inadvertently lower response rates.	
Consistency in results	Unable to assess; no measure of consistency is reported
Precision in results	Unable to assess; no measure of precision is reported
Directness of results	Direct

O'Shea LE, Mitchell AE, Picchioni MM, Dickens GL

Moderators of the predictive efficacy of the Historical, Clinical and Risk Management-20 for aggression in psychiatric facilities: Systematic review and meta-analysis

Aggression and Violent Behavior 2013; 18: 255-270

View review abstract online

Comparison	**Efficacy of the Historical, Clinical and Risk Management-20 scale (HCR-20) for predicting violence in psychiatric facilities.** **Note: the HCR-20 comprises 20 items; the Historical Scale (H10)**

<table>
<tr><td></td><td><b>contains ten items that are thought to be relatively static, and reflect the individual's psychosocial adjustment and history of violence; the Clinical Scale (C5) includes five dynamic risk factors reflecting the individual's current or recent mental health-related functioning; and the Risk Management Scale (R5) includes five dynamic risk factors that reflect professional opinions regarding the individual's ability to adjust to the institution or community. There is also a final summary judgment, which is an aggregate of the Historical and Clinical scales (HC15).</b></td></tr>
<tr><td><b>Summary of evidence</b></td><td><b>Moderate to high quality evidence (large samples, mostly precise, direct, some inconsistency) suggests good predictive validity of the Historical, Clinical and Risk Management-20 scale (HCR-20) for predicting violence in psychiatric facilities.</b>
<b>The best predictive efficacy was for samples containing higher proportions of patients with schizophrenia, women, and Caucasians.</b></td></tr>
<tr><td colspan="2"><b>Predictive validity</b></td></tr>
<tr><td colspan="2"><i>Significant, medium to large effect of predictive validity for all scales of the HCR-20 total;</i>
Any inpatient aggression: 18 studies, N = 1502, d = 0.654, 95%CI 0.436 to 0.873, $p < 0.001$, Q_w = 305.79, $p < 0.001$
Verbal aggression: 2 studies, N = 186, d = 0.932, 95%CI 0.414 to 1.45, $p < 0.001$, Q_w = 11.916, $p < 0.001$
Any physical aggression: 13 studies, N = 1271, d = 0.604, 95%CI 0.336 to 0.871, $p < 0.001$, Q_w = 276.676, $p < 0.001$
Physical to others: 10 studies, N = 1000, d = 0.421, 95%CI 0.171 to 0.673, $p < 0.01$, Q_w = 140.923, $p < 0.001$
Physical to objects: 2 studies, N = 164, d = 0.758, 95%CI -0.008 to 1.524, $p > 0.05$, Q_w = 19.688, $p < 0.001$
<i>Significant, medium effect of predictive validity of the H10 for any inpatient aggression and any physical aggression to other people. No significant effect for verbal aggression or aggression to objects;</i>
Any inpatient aggression: 20 studies, N = 1691, d = 0.423, 95%CI 0.266 to 0.58, $p < 0.001$, Q_w = 193.731, $p < 0.001$
Verbal aggression :4 studies, N = 186, d = 0.295, 95%CI -0.407 to 0.996, $p > 0.05$, Q_w = 92.380, $p < 0.001$
Any physical aggression: 15 studies, N = 1460, d = 0.375, 95%CI 0.200 to 0.551, $p < 0.001$, Q_w = 155.296, $p < 0.001$
Physical to others: 9 studies, N = 827, d = 0.299, 95%CI 0.076 to 0.522, $p < 0.01$, Q_w = 77.228, p <</td></tr>
</table>

0.001

Physical to objects: 5 studies, N = 267, *d* = 0.303, 95%CI -0.276 to 0.881 1.025, *p* > 0.05, Q_w = 82.974, *p* < 0.001

Significant, medium to large effect of predictive validity for all scales of the C5;

Any inpatient aggression: 21 studies, N = 1835, *d* = 0.743, 95%CI 0.633 to 0.854, *p* < 0.001, Q_w = 110.711, *p* < 0.001

Any physical aggression: 16 studies, N = 1145, *d* = 0.739 95%CI 0.592 to 0.885, *p* < 0.001, Q_w = 127.211, *p* < 0.001

Verbal aggression: 4 studies, N = 264, *d* = 0.970, 95%CI 0.809 to 1.132, *p* < 0.001, Q_w = 4.914, *p* > 0.05

Physical to others: 8 studies, N = 802, *d* = 0.714, 95%CI 0.545 to 0.883, *p* < 0.001, Q_w = 37.485, *p* < 0.001

Physical to objects: 4 studies, N = 242, *d* = 0.877, 95%CI 0.618 to 1.135, *p* < 0.001, Q_w = 11.034, *p* < 0.05

Significant, medium to large effect of predictive validity for all scales of the R5;

Any inpatient aggression: 14 studies, N = 1211, *d* = 0.602, 95%CI 0.428 to 0.776, *p* < 0.001, Q_w = 116.220, *p* < 0.001

Any physical aggression: 10 studies, N = 1061, *d* = 0.618, 95%CI 0.390 to 0.846, *p* < 0.001, Q_w = 123.924, *p* > 0.05

Verbal aggression: 2 studies, N = 186, *d* = 0.977, 95%CI 0.802 to 1.153, *p* < 0.001, Q_w = 1.426, *p* > 0.05

Physical to others: 6 studies, N = 724, *d* = 0.539, 95%CI 0.281 to 0.797, *p* < 0.001, Q_w = 58.772, *p* < 0.001

Physical to objects: 2 studies, N = 164, *d* = 0.832, 95%CI 0.375 to 1.289, *p* < 0.001, Q_w = 7.021 *p* < 0.01

Significant, medium effect of predictive validity of the HC15 for any inpatient aggression and any physical aggression. No significant effect for verbal aggression or aggression to other people or objects;

HC15

Any inpatient aggression: 5 studies, N = 440, *d* = 0.545, 95%CI 0.208 to 0.882, *p* < 0.01, Q_w = 47.440, *p* < 0.001

Any physical aggression: 5 studies, N = 440, *d* = 0.472, 95%CI 0.149 to 0.765, *p* < 0.01, Q_w = 43.633, *p* < 0.001

Verbal aggression: 2 studies, N = 78, *d* = 0.484, 95%CI -0.797 to 1.765, *p* > 0.05, Q_w = 32.763, *p* < 0.001

Physical to others: 2 studies, N = 78, *d* = 0.727, 95%CI- 0.271 to 1.726, *p* > 0.05, Q_w = 19.915, *p* < 0.001

Physical to objects: 2 studies, N = 78, *d* = 0.509, 95%CI -1.137 to 2.154, *p* > 0.05, Q_w = 54.068, *p* <

0.001 Moderator analyses showed that for the HCR-20 total, H10 and R5, larger effect sizes were obtained from studies with a larger proportion of patients with a diagnosis of schizophrenia. For the HC15, smaller effect sizes were obtained from studies with a larger proportion of males in their sample, and larger effect sizes obtained from studies with a higher risk of bias. For H10, larger effect sizes were associated with studies containing a larger proportion of Caucasian patients. The amount of variability not explained by each of these moderators was not significant. Authors state that 4 studies had a low risk of bias, 7 studies had a high risk of bias and 9 studies the risk of bias was unclear. No risk of publication bias was detected.	
Consistency in results	Mostly inconsistent for overall analyses, consistent for moderator analyses
Precision in results	Mostly precise
Directness of results	Direct

Papaioannou D, Brazier J, Parry G

How Valid and Responsive Are Generic Health Status Measures, such as EQ-5D and SF-36, in Schizophrenia? A Systematic Review

Value in Health 2011; 14: 907-920

View review abstract online

Comparison	**Assessment of the construct validity and responsiveness of two generic health-related quality of life profile measures; the short form health surveys (SF-36 and SF-12), and two preference-based health-related quality of life measures; short form health survey (SF-6D) and EuroQol-5D (EQ-5D) in people with schizophrenia.**
Summary of evidence	**Moderate to low quality evidence (unclear sample size, direct, unable to assess consistency or precision) suggests good 'known groups' validity for the SF-36, but inconsistent convergent validity and poor responsiveness.** **Low quality evidence (mostly single studies with unclear sample sizes) is unable to determine known groups validity or responsiveness of the EQ-5D, and psychometric properties of the SF-12 or SF-6D.**

EQ-5D
Convergent validity Symptoms 3 studies reported medium to strong correlations with the PANSS, SCL-90R, CGI-S or BPRS. 2 studies reported weak or no correlations with the PANSS. 1 study reported medium to strong correlations with depression or anxiety measures. Functioning 1 study reported no correlations with the GAF or SOFAS, but 2 studies reported medium to strong correlations with these measures. 1 study reported medium to strong correlations with the HoNOS, and weak to medium correlations with the GARF. Quality of Life 1 study reported medium correlations with the S-QoL, but 1 study reported no correlations with the QLS. **Known groups validity** 1 study reported significant differences in EQ-5D index scores between individuals defined as “severe” or “less severe” on the PANSS, HAM-D and GAF. **Responsiveness** 1 study reported weak correlations with changes > 25%, but not < 25% on the BPRS. 1 study reported significant correlations with the PANSS positive, AHRS and the GSDS, but not with any other measures of symptom severity. **Distribution properties** 2 studies reported that the EQ-5D was normally distributed, but 1 study reported a moderate ceiling effect (where 21% of participants achieved maximum score).
SF-36
Convergent validity Symptoms 5 studies reported weak or no correlations with symptom measures (various measures). 2 studies reported medium correlations with PANSS scores, and 1 study reported medium correlations with BPRS scores. 2 studies reported weak correlations with depressive symptoms measured by MADRS or CDS scores, and 1 study reported medium correlations with the MADRS or the CDSS. Functioning 1 study reported weak to medium correlations with GAF scores, and 1 study reported strong correlations with GAF scores. 2 studies reported strong correlations with the SOFAS. Quality of Life 1 study reported very weak correlations with the QLS, but 2 studies reported moderate to very

strong correlations with World Health Organization quality of life instruments. **Known groups validity** Authors report that nine of 11 studies found statistically significant differences in results between individuals with schizophrenia and normative values. **Responsiveness** 4 studies reported weak or no correlations with changes over time on the PANSS. 1 study reported weak correlations with changes on the PANSS positive scale and the MADRS. 1 study reported a weak correlation with changes on the CDSS and the ESRS. 2 studies reported no correlations for improved/remitted, or not improved/non-remitted, apart from a weak correlation with improved social functioning (reported in 1 study). **Distribution properties** 1 study reported that scores on the SF-36 were normally distributed with no evidence of floor or ceiling effects.	
SF-12	
Known groups validity 1 study reported that individuals with psychosis were significantly more likely to report disability on the SF-12 than individuals with no mental health disorder.	
SF-6D	
Convergent validity and responsiveness Symptoms 1 study reported medium correlations with the BPRS, and when changes occurred on the BPRS (> 25%), changes in the SF-6D were correlated weakly. **Distribution properties** The same study reported that scores on the SF-6D were found to be normally distributed with no evidence of floor or ceiling effects.	
Consistency in results	Unable to assess; no measure of consistency is reported
Precision in results	Unable to assess; no measure of precision is reported
Directness of results	Direct

Preston E, Hansen L

A systematic review of suicide rating scales in schizophrenia

Crisis: Journal of Crisis Intervention & Suicide 2005; 26(4): 170-80 View review abstract online	
Comparison	**Description of tools for assessing suicide risk in people with schizophrenia.**
Summary of evidence	**Moderate to low quality evidence (mostly small samples, direct, unable to assess consistency or precision) is unclear as to the most effective assessment scaled for suicidality in people with schizophrenia.**
Suicidal risk scales	
Five scales were identified that aim to predict suicidality in people with schizophrenia. Stephens' Scale for Suicide Risk in Schizophrenia: N = 1212 inpatients. High score was insufficiently specific to predict suicide risk but may be a useful warning for suicide potential. Note that the scale was constructed their own scale based on risk factors found in their study. Schizophrenia Suicide Risk Scale: N = 69, inter-rater reliability kappa = 0.79 (SD 0.30). Authors found this scale was of most benefit for patients at very low or very high risk of suicide but lacked sensitivity for detecting people at medium risk. The items with the strongest predictive power were 'communicated suicide plans', 'suicide attempts', 'loss of job', 'observed depression', and 'suicide plans'. Scale for Suicide Ideation: N = 105, high correlations (> 0.90) between patient self-report and clinician rated versions. The sample included other diagnoses so the utility for schizophrenia is unclear. Beck Scale for Suicide Ideation: self-report version of the SSI, N = 142 inpatients. High correlations reported between BSI score and previous suicide attempts. The sample also included affective psychoses so the utility for schizophrenia is unclear. InterSePT Scale for Suicidal Thinking: Study 1: N = 22, mean inter-rater reliability kappa = 0.90. Study 2: N = 980, found the ISST total score was associated with PANSS scores and measures of substance use. Authors suggest this scale may currently present the most useful option for assessing suicide risk.	
Consistency in results	Unable to assess; no measure of consistency is reported
Precision in results	Unable to assess; no measure of precision is reported
Directness of results	Direct

Register-Brown K, Hong LE

Reliability and validity of methods for measuring the duration of untreated

psychosis: A quantitative review and meta-analysis **Schizophrenia Research 160; 2014: 20-26** View review abstract online	
Comparison	**Reliability and validity of assessment tools that measure the duration of untreated psychosis, psychosis onset and treatment onset.**
Summary of evidence	**Moderate quality evidence (medium to large samples, direct, inconsistent, unable to assess precision) suggests good inter-rater reliability and small predictive validity for tools assessing duration of untreated psychosis, psychosis onset and treatment onset.**
Inter-rater reliability	
Authors state that all scales had good inter-rater reliability; Clinical Interview: 55 studies, N = 10,089, DUP ICC = 0.7 to 1.0 Chart Review: 6 studies, N = 497, DUP ICC = 0.73 Beiser Scale: 11 studies, N = 786, DUP ICC = 0.79 to 0.98, Psychosis onset ICC = 0.94 to 0.98, Treatment onset ICC = 0.95 Comprehensive Assessment of Symptoms and History: 4 studies, N = 337, DUP ICC = 0.87 to 1.00, Psychosis onset ICC = 0.96, Treatment onset ICC = 0.96 to 1.00 Circumstances of Onset and Relapse Schedule: 7 studies, N = 259, DUP ICC= 0.71 to 0.98 Interview for the Retrospective Assessment of the Onset of Schizophrenia: 11 studies, N = 1089, DUP k = 0.6 to 0.95, Psychosis onset PA = 77%, Treatment onset PA = 80 to 100% Nottingham Onset Schedule: 2 studies, N = 1740, DUP ICC = 0.95 to 0.99, Psychosis onset PA = 70% Positive and Negative Syndrome Scale for Schizophrenia (modified): 18 studies, N = 1969, DUP ICC = 0.9 to 0.99 Psychiatric and Personal History Schedule: 4 studies, N = 277, DUP ICC = 0.90 Royal Park Multidiagnostic Instrument for Psychosis: 6 studies, N = 661, DUP k = 0.79, Psychosis onset k = 0.79 Symptom Onset in Schizophrenia Inventory: 7 studies, N = 937, DUP ICC = 0.99, Psychosis onset ICC = 1.0	
Predictive validity	
Authors report small effect sizes overall, and that no instrument had clearly larger effect sizes across different categories of outcomes or when all outcomes were grouped together.	

All scales combined had significant predictive value for;

All outcomes combined: 132 studies, z = 0.18, $p < 0.001$

Treatment adherence: 6 studies, z = 0.14, $p < 0.001$

Overall functioning: 49 studies, z = 0.22, $p < 0.001$

Imaging outcomes: 14 studies, z = 0.25, $p < 0.001$

Negative symptoms: 32 studies, z = 0.21, $p < 0.001$

Positive symptoms: 27 studies, z = 0.22, $p < 0.001$

Neurocognition: 19 studies, z = 0.20, $p < 0.001$

Relapse risk: 29 studies, z = 0.21, $p < 0.001$

Suicidality/ violence: 15 studies, z = 0.084, $p < 0.001$

Clinical interview had significant predictive value for;

All outcomes combined: 55 studies, z = 0.17, $p < 0.001$

Treatment adherence: 3 studies, z = 0.15, $p < 0.05$

Overall functioning: 18 studies, z = 0.21, $p < 0.001$

Imaging outcomes: 8 studies, z = 0.32, $p < 0.001$

Negative symptoms: 11 studies, z = 0.19, $p < 0.001$

Positive symptoms: 9 studies, z = 0.24, $p < 0.001$

Neurocognition: 8 studies, z = 0.29, $p < 0.001$

Relapse risk: 12 studies, z = 0.22, $p < 0.001$

But not for:

Suicidality/violence: 5 studies, z = 0.07, $p > 0.05$

Chart review had significant predictive value for;

All outcomes combined: 6 studies, z = 0.32, $p < 0.001$

Relapse risk: 2 studies, z = 0.37, $p < 0.05$

But not for;

Imaging outcomes: 1 study, z = 0.27, $p > 0.05$

Suicidality/violence: 2 studies, z = 0.03, $p > 0.05$

Overall functioning: 3 studies, z = 0.14, $p > 0.05$

Basel Interview had no significant predictive value for;

Neurocognition: 1 study, z = 0.19, $p > 0.05$

Beiser Scale had significant predictive value for;

All outcomes combined: 11 studies, z = 0.20, $p < 0.001$

Overall functioning: 5 studies, z = 0.30, $p < 0.05$

Negative symptoms: 2 studies, z = 0.33, $p < 0.001$

Positive symptoms: 3 studies, z = 0.28, $p < 0.001$

Neurocognition: 1 study, z = 0.44, $p < 0.05$

Suicidality/ violence: 2 studies, z = 0.16, $p < 0.05$

But not for;

Treatment adherence: 2 studies, z = 0.05, $p > 0.05$

Relapse risk: 2 studies, z = 0.02, $p > 0.05$

Comprehensive Assessment of Symptoms and History had significant predictive value for;

All outcomes combined: 4 studies, z = 0.12, $p < 0.05$

Overall functioning: 2 studies, z = 0.14, $p < 0.05$

Imaging outcomes: 2 studies, z = 0.14, $p < 0.05$

But not for;

Negative symptoms: 1 study, z = 0.02, $p > 0.05$

Positive symptoms: 1 study, z = 0.12, $p > 0.05$

Neurocognition: 1 study, z = 0.15, $p > 0.05$

Circumstances of Onset and Relapse Schedule had significant predictive value for;

All outcomes combined: 7 studies, z = 0.006, $p < 0.05$

Overall functioning: 2 studies, z = 0.19, $p < 0.001$

Imaging outcomes: 1 study, z = 0.22, $p < 0.05$

Negative symptoms: 3 studies, z = 0.18, $p < 0.05$

Positive symptoms: 2 studies, z = 0.22, $p < 0.001$

But not for:

Neurocognition: 2 studies, z = 0.19, $p > 0.05$

Relapse risk: 2 studies, z = 0.11, $p > 0.05$

Interview for the Retrospective Assessment of the Onset of Schizophrenia had significant predictive value for;

All outcomes combined: 11 studies, z = 0.17, $p < 0.001$

Overall functioning: 5 studies, z = 0.17, $p < 0.001$

Negative symptoms: 5 studies, z = 0.10, $p < 0.05$

Positive symptoms: 4 studies, z = 0.15, $p < 0.05$

Neurocognition: 2 studies, z = 0.26, $p < 0.05$

But not for;

Relapse risk: 4 studies, z = 0.14, $p > 0.05$

Nottingham Onset Schedule had significant predictive value for;

Imaging outcomes: 1 study, z = 0.68, $p < 0.001$

But not for;

Suicidality/ violence: 1 study, z = -0.02, $p > 0.05$

Overall functioning: 2 studies, z = 0.12, $p > 0.05$

Positive and Negative Syndrome Scale for Schizophrenia (modified) had significant predictive value for;

All outcomes combined: 18 studies, z = 0.16, $p < 0.001$

Treatment adherence: 1 study, z = 0.19, $p < 0.05$

Overall functioning: 10 studies, z = 0.20, $p < 0.001$

Negative symptoms: 7 studies, z = 0.20, $p < 0.05$

Relapse risk: 5 studies, z = 0.15, $p < 0.05$

Suicidality/ violence: 3 studies, z = 0.19, $p < 0.001$

But not for;

Positive symptoms: 6 studies, z = 0.11, $p > 0.05$

Neurocognition: 1 study, z = 0.10, $p > 0.05$

Psychiatric and Personal History Schedule had significant predictive value for;

All outcomes combined: 4 studies, z = 0.23, $p < 0.05$

Imaging outcomes: 1 study, z = 0.35, $p < 0.05$

But not for;

Neurocognition: 1 study, z = 0.03, $p > 0.05$

Overall functioning: 2 studies, z = 0.45, $p > 0.05$

Royal Park Multi-diagnostic Instrument for Psychosis had significant predictive value for;

All outcomes combined: 6 studies, z = 0.20, $p < 0.001$

Overall functioning: 3 studies, z = 0.33, $p < 0.001$

Negative symptoms: 2 studies, z = 0.29, $p < 0.001$

Positive symptoms: 2 studies, z = 0.31, $p < 0.05$

Neurocognition: 1 study, z = 0.38, $p < 0.001$

Relapse risk: 1 study, z = 0.33, $p < 0.001$

But not for:

Suicidality/ violence: 1 study, z = 0.02, $p > 0.05$

Symptom Onset in Schizophrenia Inventory had significant predictive value for;

All outcomes combined: 7 studies, z = 0.16, $p < 0.001$

Negative symptoms: 1 study, z = 0.27, *p* < 0.001 Relapse risk: 1 study, z = 0.28, *p* < 0.001 *But not for;* Overall functioning: 2 studies, z = 0.16, *p* > 0.05 Neurocognition: 1 study, z = -0.09, *p* > 0.05 Suicidality/ violence: 1 study, z = 0.01, *p* > 0.05 Additional meta-analyses of DUP measured by any specialized instrument vs. generic clinical interviews revealed no difference in effect size on any outcome. Authors report no evidence of publication bias.	
Consistency in results	Unable to assess inter-rater reliability, authors report moderate to high heterogeneity for predictive validity
Precision in results	Unable to assess; no measure of precision is reported
Directness of results	Direct

Samara MT, Leucht C, Leeflang MM, Anghelescu IG Chung YC, Crespo-Facorro B, Elkis H, Hatta K, Giegling I, Kane JM, Kayo M, Lambert M, Lin CH, Möller HJ, Pelayo-Terán JM, Riedel M, Rujescu D, Schimmelmann BG, Serretti A, Correll CU, Leucht S **Early Improvement As a Predictor of Later Response to Antipsychotics in Schizophrenia: A Diagnostic Test Review** **American Journal of Psychiatry 2015; 172: 617-629** View review abstract online	
Comparison	**Predictive value of scales measuring oral antipsychotic response at 2 weeks for response at the end of the study (4 to 12 weeks). Dosage was within target dose range, but lower doses were acceptable in studies of first-episode psychosis patients or adolescents.**
Summary of evidence	**Moderate to high quality evidence (large sample, direct, appears precise, unable to assess consistency) suggests good predictive value of a less than 20% reduction on BPRS or PANSS scores at 2 weeks after baseline for non-response at the end of the study (4-12 weeks). Higher specificity was associated with shorter trial duration, higher baseline illness severity, and**

	shorter illness duration.
BPRS and PANSS predictive validity	
Good predictive value of non-response at endpoint with a < 20% PANSS or BPRS score reduction at week 2; 34 studies, N = 9,460 Specificity = 86%, 95%CI 0.82 to 0.89 Sensitivity = 63%, 95%CI 0.59 to 0.66 Positive predictive value = 90%, 95%CI 0.86 to 0.91 Negative predictive value = 53% 95%CI 0.49 to 0.61 Authors report that higher specificity was associated with shorter trial duration, higher baseline illness severity, and shorter illness duration.	
Consistency in results	Unable to assess; no measure of consistency is reported
Precision in results	Appears precise
Directness of results	Direct

Shafer A

Meta-analysis of the brief psychiatric rating scale factor structure

Psychological Assessment 2005; 17(3): 324-35

View review abstract online

Comparison	**Description of the Brief Psychiatric Rating Scale factor structure.**
Summary of evidence	**Moderate quality evidence (large sample size, direct, unable to assess consistency or precision) suggests that the BPRS can be factored into several discrete components, comprising positive, negative, and affective symptoms, resistance (hostility) and activation (excitement).**
BPRS item structure	
Meta-analysis was performed on 26 studies (N = 17,620) that conducted factor analyses on the BPRS items (18- or 24-item versions) Authors reported that the 18-item BPRS can be effectively structured into a four- or five-component	

solution, where the four-component solution accounts for 76% of the variance, and the five-factor solution accounts for 88% of the variance. Both four- and five-factor solutions demonstrate a clear affective component that includes items for depression, anxiety, guilt, and somatic concern. A positive symptom component comprises unusual thought content, hallucinations, grandiosity, and conceptual disorganisation. A negative symptom component comprises blunted affect, emotional withdrawal, disorientation, and motor retardation. A resistance component comprises hostility, uncooperativeness, and suspiciousness. The five-component solution additionally included an activation component, defined by excitement, tension, and mannerisms/posturing.	
Consistency in results	Unable to assess; no measure of consistency is reported
Precision in results	Unable to assess; no measure of precision is reported
Directness of results	Direct assessment

Shafer A, Dazzi F

Meta-analysis of the positive and Negative Syndrome Scale (PANSS) factor structure

Journal of Psychiatric Research 2019; 115: 113-20

View review abstract online

Comparison	**Description of the PANSS factor structure.**
Summary of evidence	**Moderate quality evidence (large sample size, direct, unable to assess consistency or precision) suggests that the PANSS can be factored into several discrete components, comprising positive, negative, disorganised, and affective symptoms, resistance (hostility) and activation (excitement).**
PANSS item structure	
Meta-analysis was performed on 45 studies (N = 22,812) that conducted factor analyses on the PANSS items *The factors and the items defining them were;* Positive symptoms: delusions, unusual thought content, hallucinatory behaviour, suspiciousness and persecution, and grandiosity Negative symptoms: emotional withdrawal, blunted affect, passive apathetic social withdrawal, lack of spontaneity, poor rapport, motor retardation, active social avoidance Disorganisation/cognitive symptoms: conceptual disorganisation, poor attention, eifficulty in abstract	

thinking, disturbance of volition, stereotyped thinking, mannerisms/posturing, preoccupation, disorientation Depression/anxiety symptoms: anxiety, depression, guilt feelings, tension, somatic concern Resistance/excitement: hostility, poor impulse control, excitement, uncooperativeness Compared to the BPRS, the PANSS was distinguished by a larger number of items to clearly define negative symptoms and a disorganisation factor.	
Consistency in results	Unable to assess; no measure of consistency is reported
Precision in results	Unable to assess; no measure of precision is reported
Directness of results	Direct assessment

Singh JP, Serper M, Reinharth J, Faz S

Structured Assessment of Violence Risk in Schizophrenia and Other Psychiatric Disorders: A Systematic Review of the Validity, Reliability, and Item Content of 10 Available Instruments

Schizophrenia Bulletin 2011; 37(5): 899-912

View review abstract online

Comparison	**Psychometric properties of scales measuring violence risk in community outpatients with schizophrenia or psychosis.**
Summary of evidence	**Moderate to low quality evidence (small sample, appears precise, unable to assess consistency, direct) suggests poor predictive validity of the Violence Risk Appraisal Guide for assessing violence in people with schizophrenia living in the community.**
Predictive validity	
Violence Risk Appraisal Guide (VRAG) 2 studies (N = 165) reported AUC data for patients with schizophrenia Mean length of follow up = 64.4 months, median AUC = 0.69, interquartile range = 0.60 to 0.77 *Authors conclude that there is currently little direct evidence for violence risk assessment tools' utility in individuals with schizophrenia.*	
Consistency in results	Unable to assess; no measure of consistency is reported

Precision in results	Appears precise
Directness of results	Direct

Smith EL, Garety PA, Harding H, Hardy A

Are there reliable and valid measures of anxiety for people with psychosis? A systematic review of psychometric properties

Psychology and psychotherapy 2021; 94(1): 173-98

View review abstract online

Comparison	**Psychometric properties of scales measuring anxiety in people with schizophrenia or psychosis.**
Summary of evidence	**Moderate quality evidence (large sample, unable to assess consistency or precision, direct) suggests the Beck Anxiety Index, Depression Anxiety Stress Scale or Scale of Anxiety Evaluation in Schizophrenia for general screening, and the DSM-based Generalised Anxiety Disorder Symptoms Severity Scale, Liebowitz Social Anxiety Scale, Obsessive-Compulsive Inventory, Psychological Stress Index, Perseverative Thinking Questionnaire, and Yale-Brown Obsessive Compulsive Scale to assess anxiety symptoms.**

Psychometric properties

11 studies, N = 1,453

The Scale of Anxiety Evaluation in Schizophrenia demonstrated consistently good psychometric properties.

The Beck Anxiety Index, Depression Anxiety Stress Scale, DSM-based Generalised Anxiety Disorder Symptoms Severity Scale, Liebowitz Social Anxiety Scale, Obsessive-Compulsive Inventory, Psychological Stress Index, Perseverative Thinking Questionnaire, and Yale-Brown Obsessive Compulsive Scale demonstrated adequate reliability and/or validity.

Authors report methodological quality was largely poor according to the requirements of the COSMIN checklist.

Authors recommend the Beck Anxiety Index, Depression Anxiety Stress Scale or Scale of Anxiety Evaluation in Schizophrenia for general screening, and the DSM-based Generalised Anxiety Disorder Symptoms Severity Scale, Liebowitz Social Anxiety Scale, Obsessive-Compulsive Inventory, Psychological Stress Index, Perseverative Thinking Questionnaire, and Yale-Brown Obsessive Compulsive Scale to assess anxiety symptoms.

Consistency in results	Unable to assess; no measure of consistency is reported.
Precision in results	Unable to assess; no measure of precision is reported.
Directness of results	Direct

<table>
<tr><td colspan="2"><em>Yeh YC, Lin CY, Li PC, Hung CF, Cheng CH, Kuo MH, Chen KL</em>
<strong>A systematic review of the current measures of theory of mind in adults with Schizophrenia</strong>
<strong>International Journal of Environmental Research and Public Health 2021; 18(13): 7172</strong>
View review abstract online</td></tr>
<tr><td><strong>Comparison</strong></td><td><strong>Description of tools for assessing theory of mind.</strong></td></tr>
<tr><td><strong>Summary of evidence</strong></td><td><strong>Moderate to low quality evidence (unclear sample size, unable to assess consistency or precision, direct) suggests good internal consistencies for the Visual Jokes task, Faux Pas task, Reading the Mind in the Eyes Test, and the Moving Shapes task. The Hinting task and False Belief picture Sequencing showed moderate internal consistencies.</strong>
<strong>Good test-retest reliabilities were reported for the Hinting task and the Faux Pas task. The Story test had moderate test-retest reliability. The False Belief stories task and the second-order False Belief stories task had poor reliability. The Reading the Mind in the Eyes Test had inconsistent reliability measures.</strong></td></tr>
<tr><td colspan="2"><strong>Internal consistency</strong></td></tr>
<tr><td colspan="2">117 studies, N not reported
Good internal consistencies were reported for the Visual Jokes task (α = 0.83), Faux Pas task (α = 0.82), Reading the Mind in the Eyes Test (α = 0.73), and Moving Shapes task (α = 0.80-0.84). The Hinting task (α = 0.57) and False Belief picture Sequencing (α = 0.54) showed moderate internal consistencies.</td></tr>
<tr><td colspan="2"><strong>Test-retest reliability</strong></td></tr>
<tr><td colspan="2">117 studies, N not reported
Good test-retest reliabilities were reported for the Hinting task (ICC = 0.78), and the Faux Pas task (ICC = 0.76). The Story test had moderate test-retest reliability (ICC = 0.50). The False Belief stories task (ICC = 0.31) and the second-order False Belief stories task (ICC = 0.31) had poor</td></tr>
</table>

reliability. The Reading the Mind in the Eyes Test had inconsistent reliability measures (ICC = 0.24, *r* = 0.78).	
Consistency in results	Unable to assess; no measure of consistency is reported.
Precision in results	Unable to assess; no measure of precision is reported.
Directness of results	Direct

Zeller SL, Rhoades RW

Systematic reviews of assessment measures and pharmacologic treatments for agitation

Clinical Therapeutics, 2010. 32(3): p. 403-425

View review abstract online

Comparison	**Description of tools for assessing agitation and risk of aggression/violence.**
Summary of evidence	**Moderate to low quality evidence (unclear sample size, unable to assess consistency or precision, direct) suggests the most effective assessment scales for predicting aggression/violence are the Historical, Clinical and Risk Manaagement-20, the McNiel-Binder Violence Screening Checklist, and the Brøset Violence Checklist.**

Agitation scales

Thirteen scales were identified that assess the severity of agitation and aim to predict possible aggression/violence, and can be applied to people with schizophrenia:

Aggressive Behavior Scale; Agitated Behavior Scale; Brief Agitation Rating Scale; Brøset Violence Checklist; Clinical Global Impression Scale for Aggression; Cohen-Mansfield Agitation Inventory; Historical, Clinical, and Risk Management–20 Violence Risk Assessment Scheme; McNiel-Binder Violence Screening Checklist; Neurobehavioral Rating Scale–Revised; Overt Aggression Scale; Overt Agitation Severity Scale; Positive and Negative Syndrome Scale–Excited Component; and the Ryden Aggression Scale.

Authors report that these scales vary widely in suitability for application outside research settings.

Only three scales reported acceptable accuracy for predicting aggression/violence: the Historical, Clinical and Risk Manaagement-20; the McNiel-Binder Violence Screening Checklist; and the Brøset Violence Checklist. The PANSS-EC scale has also been used in practice to assess patients' need for psychotropic medication.

Consistency in results	Unable to assess; no measure of consistency is reported.
Precision in results	Unable to assess; no measure of precision is reported.
Directness of results	Direct

Explanation of acronyms

α = Cronbach's alpha, AHRS = auditory hallucinations rating scale, BDI = Beck Depression Inventory, BPRS = Brief Psychiatric Rating Scale, BSABS = Bonn Scale for the Assessment of Basic Symptoms, BSI = Beck Scale for Suicide Ideation, CAARMS = Comprehensive Assessment of At-Risk Mental States, CDSS = Calgary Depression Scale for Schizophrenia, CGI-S = Clinical Global Impression – Severity scale, ERIraos = Early Recognition Inventory, EASE = Examination of Anomalies in Self-experience, EQ-5D = EuroQol-5D, ESRS = Extrapyramidal Symptom Rating Scale, GAF = Global Assessment of Functioning, z = Fisher's z distribution, GARF = Global Assessment of Relational Functioning, GSDS = Groningen social disabilities schedule, HAMD = Hamilton Rating Scale for Depression, HoNOS = Health of the Nation Outcome Scales, ICC = intraclass correlation, k = Cohen's kappa coefficient, MADRS = Montgomery Asberg Depression Rating Scale, N = number of participants, NPV = negative predictive value - the proportion of patients with negative test results who are correctly diagnose, PA = pairwise agreement, PANSS = Positive and negative syndrome scale, PANSS-D = depression scale, PANSS-EC = excited scale, PPV = positive predictive value - proportion of patients with positive test results who are correctly diagnosed, PQ, = Prodromal Questionnaire, PRODscreen = Prodromal screening test, QLS = Quality of Life Scale, Q_B = test for heterogeneity between groups of studies, Q_w = test for heterogeneity with groups of studies, SCL-90R – Symptom Checklist-90-Revised, SF-36/SF-12/ SF-6D = Short form health surveys, SIPS = Structured Interview of Prodromal Syndromes, SOFAS = Social and Occupational Functioning Assessment Scale, SPI-A = Schizophrenia Prediction Instrument – Adult version, S-QoL = Schizophrenia Quality of Life Questionnaire, SSI = Scale for Suicide Ideation, UHR = Ultra High Risk for psychosis, Y-PARQ = Youth Psychosis At Risk Questionnaire

Outcome assessment tools

Explanation of technical terms

* Bias has the potential to affect reviews of both RCT and observational studies. Forms of bias include; reporting bias – selective reporting of results; publication bias - trials that are not formally published tend to show less effect than published trials, further if there are statistically significant differences between groups in a trial, these trial results tend to get published before those of trials without significant differences; language bias – only including English language reports; funding bias - source of funding for the primary research with selective reporting of results within primary studies; outcome variable selection bias; database bias - including reports from some databases and not others; citation bias - preferential citation of authors. Trials can also be subject to bias when evaluators are not blind to treatment condition and selection bias of participants if trial samples are small[21].

† Different effect measures are reported by different reviews.

Prevalence refers to how many existing cases there are at a particular point in time. Incidence refers to how many new cases there are per population in a specified time period. Incidence is usually reported as the number of new cases per 100,000 people per year. Alternatively some studies present the number of new cases that have accumulated over several years against a person-years denominator. This denominator is the sum of individual units of time that the persons in the population are at risk of becoming a case. It takes into account the size of the underlying population sample and its age structure over the duration of observation.

Reliability and validity refers to how accurate the instrument is. Sensitivity is the proportion of actual positives that are correctly identified (100% sensitivity = correct identification of all actual positives) and specificity is the proportion of negatives that are correctly identified (100% specificity = not identifying anyone as positive if they are truly not). A receiver operating characteristic (ROC) curve represents sensitivity/specificity pairs corresponding to different cut-off values. A guide for interpreting the area under the curve (AUC) statistic is; 0.90 to 1.00 = excellent, 0.80 to 0.90 = good, 0.70 to 0.80 = fair, 0.60 to 0.70 = poor, and 0.50 to 0.60 = fail.

Weighted mean difference scores refer to mean differences between treatment and comparison groups after treatment (or occasionally pre to post treatment) and in a randomized trial there is an assumption that both groups are comparable on this measure prior to treatment. Standardized mean differences are divided by the pooled standard deviation (or the standard deviation of one group when groups are homogenous) that allows results from different scales to be combined and compared. Each study's mean difference is then given a weighting depending on the size of the sample and the variability in the data. 0.2 represents a small effect, 0.5 a moderate effect, and 0.8 and over represents a large effect[21].

Odds ratio (OR) or relative risk (RR) refers to the probability of a reduction (< 1) or an increase (> 1) in a particular outcome in a treatment group, or a group exposed to a risk factor, relative to the comparison group. For example, a RR of 0.75 translates to a reduction in risk of an outcome of 25% relative to those not receiving the treatment or not exposed to the risk factor. Conversely, a RR of 1.25 translates to an increased risk of 25% relative to those not receiving treatment or not having been exposed to a risk factor. A RR or OR of 1.00 means there is no difference between groups. A medium effect is considered if RR > 2 or < 0.5 and a large effect if RR > 5 or < 0.2[22]. lnOR stands for logarithmic OR where a lnOR of 0 shows no difference between groups. Hazard ratios measure the effect of an explanatory variable on the hazard or risk of an event.

Correlation coefficients (eg, r) indicate the strength of association or relationship between variables. They can provide an indirect indication of prediction, but do not confirm causality due to possible and often unforseen confounding variables. An r of 0.10 represents a weak association, 0.25 a medium association and 0.40 and over represents a strong association. Unstandardized (*b*) regression coefficients indicate the average change in the dependent variable associated with a 1 unit change in the independent variable, statistically controlling for the other independent variables. Standardized regression coefficients represent the change being in units of standard deviations to allow comparison across different scales.

‡ Inconsistency refers to differing estimates of effect across studies (i.e. heterogeneity or variability in results) that is not explained by subgroup analyses and therefore reduces confidence in the effect estimate. I^2 is the percentage of the variability in effect estimates that is due to heterogeneity rather than sampling error (chance) - 0% to 40%: heterogeneity might not be important, 30% to 60%: may represent moderate heterogeneity, 50% to 90%: may represent considerable heterogeneity and over this is considerable heterogeneity. I^2 can be calculated from Q (chi-square) for the test of heterogeneity with the following formula[21];

$$I^2 = \left(\frac{Q - df}{Q}\right) \times 100\%$$

§ Imprecision refers to wide confidence intervals indicating a lack of confidence in the effect estimate. Based on GRADE recommendations, a result for continuous data (standardised mean differences, not weighted mean differences) is considered imprecise if the upper or lower confidence limit crosses an effect size of 0.5 in either direction, and for binary and correlation data, an effect size of 0.25. GRADE also recommends downgrading the evidence when sample size is smaller than 300 (for binary data) and 400 (for continuous data), although for some topics, these criteria should be relaxed[23].

‖ Indirectness of comparison occurs when a comparison of intervention A versus B is not available but A was compared with C and B was compared with C that allows indirect comparisons of the magnitude of effect of A versus B. Indirectness of population, comparator and/or outcome can also occur when the available evidence regarding a particular population, intervention, comparator, or outcome is not available and is therefore inferred from available evidence. These inferred treatment effect sizes are of lower quality than those gained from head-to-head comparisons of A and B.

References

1. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMAGroup (2009): Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *British Medical Journal* 151: 264-9.
2. GRADEWorkingGroup (2004): Grading quality of evidence and strength of recommendations. *British Medical Journal* 328: 1490.
3. Preston E, Hansen L (2005): A systematic review of suicide rating scales in schizophrenia. *Crisis: Journal of Crisis Intervention & Suicide* 26: 170-80.
4. Zeller SL, Rhoades RW (2010): Systematic reviews of assessment measures and pharmacologic treatments for agitation. *Clinical Therapeutics* 32: 403-25.
5. Shafer A (2005): Meta-analysis of the brief psychiatric rating scale factor structure. *Psychological Assessment* 17: 324-35.
6. Cavelti M, Kvrgic S, Beck EM, Kossowsky J, Vauth R (2012): Assessing recovery from schizophrenia as an individual process. A review of self-report instruments. *European Psychiatry: the Journal of the Association of European Psychiatrists* 27: 19-32.
7. Lako IM, Bruggeman R, Knegtering H, Wiersma D, Schoevers RA, Slooff CJ*, et al.* (2012): A systematic review of instruments to measure depressive symptoms in patients with schizophrenia. *Journal of Affective Disorders* 140: 38-47.
8. Obermeier M, Schennach-Wolff R, Meyer S, Moller HJ, Riedel M, Krause D*, et al.* (2011): Is the PANSS used correctly? A systematic review. *BMC Psychiatry* 11.
9. Singh JP, Serper M, Reinharth J, Fazel S (2011): Structured assessment of violence risk in schizophrenia and other psychiatric disorders: a systematic review of the validity, reliability, and item content of 10 available instruments. *Schizophrenia Bulletin* 37: 899-912.
10. Bakkour N, Samp J, Akhras K, El Hammi E, Soussi I, Zahra F*, et al.* (2014): Systematic review of appropriate cognitive assessment instruments used in clinical trials of schizophrenia, major depressive disorder and bipolar disorder. *Psychiatry Research* 216: 291-302.
11. Dumas R, Baumstarck K, Michel P, Lancon C, Auquier P, Boyer L (2013): Systematic review reveals heterogeneity in the use of the Scale to Assess Unawareness of Mental Disorder (SUMD). *Current Psychiatry Reports* 15: 361.
12. O'Shea LE, Mitchell AE, Picchioni MM, Dickens GL (2013): Moderators of the predictive efficacy of the Historical, Clinical and Risk Management-20 for aggression in psychiatric facilities: Systematic review and meta-analysis. *Aggression and Violent Behavior* 18: 255-70.
13. Samara MT, Leucht C, Leeflang MM, Anghelescu IG, Chung YC, Crespo-Facorro B*, et al.* (2015): Early Improvement As a Predictor of Later Response to Antipsychotics in Schizophrenia: A Diagnostic Test Review. *American Journal of Psychiatry* 172: 617-29.
14. Register-Brown K, Hong LE (2014): Reliability and validity of methods for measuring the duration of untreated psychosis: A quantitative review and meta-analysis. *Schizophrenia Research* 160: 20-6.
15. Luther L, Firmin RL, Lysaker PH, Minor KS, Salyers MP (2018): A meta-analytic review of self-reported, clinician-rated, and performance-based motivation measures in schizophrenia: Are we measuring the same "stuff"? *Clinical Psychology Review* 61: 24-37.
16. Papaioannou D, Brazier J, Parry G (2011): How valid and responsive are generic health status measures, such as EQ-5D and SF-36, in schizophrenia? A systematic review. *Value in Health* 14: 907-20.
17. Shafer A, Dazzi F (2019): Meta-analysis of the positive and Negative Syndrome Scale (PANSS) factor structure. *Journal of Psychiatric Research* 115: 113-20.
18. Erkoreka L, Ozamiz-Etxebarria N, Ruiz O, Ballesteros J (2020): Assessment of psychiatric symptomatology in bilingual psychotic patients: A systematic review and meta-analysis. *International Journal of Environmental Research and Public Health* 17(11): 1-11.
19. Smith EL, Garety PA, Harding H, Hardy A (2021): Are there reliable and valid measures of anxiety for people with psychosis? A systematic review of psychometric properties. *Psychology and psychotherapy* 94(1): 173-98.

20. Yeh YC, Lin CY, Li PC, Hung CF, Cheng CH, Kuo MH, *et al.* (2021): A systematic review of the current measures of theory of mind in adults with Schizophrenia. *International Journal of Environmental Research and Public Health* 18(13) (no pagination).

21. CochraneCollaboration (2008): Cochrane Handbook for Systematic Reviews of Interventions. Accessed 24/06/2011.

22. Rosenthal JA (1996): Qualitative Descriptors of Strength of Association and Effect Size. *Journal of Social Service Research* 21: 37-59.

23. GRADEpro (2008): [Computer program]. Jan Brozek, Andrew Oxman, Holger Schünemann. *Version 32 for Windows*