Supplementary materials

Table 1. Instrument characteristics.

Reference	Research aim(s)	PREM developed (abbreviation)	Country of origin, language	Sample studied	No. of items/dimensions	Response options	Domains/factors	Mode of administration	Time frame	Response rate (%)
Aloba et al, 2014 (43) **Anderson et, 1990 (44)**	To evaluate the psychometric characteristics of the Trust in Physician Scale among a cross-sectional sample of stable Nigerian outpatients receiving treatment for psychiatric disorders	Trust in Physician Scale (TPS)*[a]	US, English	223 adult outpatients attending the psychiatric clinics Inclusion criteria were: not actively psychotic, have affective symptoms in remission, have been diagnosed and receiving outpatient treatment for at least a year. Exclusion criteria were: comorbid chronic medical illnesses and refused	11 items/ 2 dimensions	5-point Likert	Doubt or uncertainly; Trustworthiness	Administration during an interview	NA	NA

				consent						
Atkinson et al, 2004 (45)	To develop and psychometrically evaluate a general measure of patients' satisfaction with medication, the Treatment Satisfaction Questionnaire for Medication	Treatment Satisfaction Questionnaire for Medication (TSQM)*[a]	US, English	567 patients recruited from a national longitudinal panel study of chronic illness, the NFO - World Group's CAP Inclusion criteria were: to be at least 18 years of age, able to read English, and able to complete a questionnaire on-line. Participants had to have at least one of the following illness conditions: arthritis, asthma, major depression, type I diabetes, high	14 items/ 4 dimensions	5 or 7-point Likert + yes/no format for screening questions	Side effects; Effectiveness; Convenience; Global satisfaction	Self-completion on-line	NA	67.2

				cholesterol, hypertension, migraine and psoriasis						
Baker, 1990 (46)	To develop a new questionnaire to assess patients' satisfaction with consultations together with initial tests of the questionnaire's reliability and validity	Consultation Satisfaction Questionnaire (CSQ)*[a]	UK, English	239 patients Exclusion: if they were under 16 years of age, too ill to complete the form, unable to read the form or if they had already completed any version of the questionnaire	18 items/ 4 dimensions	5-point Likert	Professional care; Depth of relationship; Perceived time; General satisfaction	Self-completion	After consultation but before departure	75.0
Barker et al, 1999 (47) **Barker et al, 1996 (48)**	To measure the reliability and validity of the Psychiatric Care Satisfaction Questionnaire (PCSQ)	Psychiatric Care Satisfaction Questionnaire (PCSQ)*	UK, English	137 and 52 inpatients from acute psychiatric unit Exclusion criteria were: patients judged by staff to be too disturbed, incoherent or distressed	18 items/ 2 dimensions	5-point Likert	General satisfaction and views of general quality services; Attitudes towards psychiatric doctors	Self-completion	NA	74.3-79.2

				and those who were unavailable						
Berghofer et al, 2011 (49)	To describe the development and to psychometrically test a new self-report tool of treatment satisfaction among people with chronic mental illnesses using community services	Evaluation of Client Services (ECS)*	US, English	184 outpatients with chronic and severe mental illnesses in community treatment settings Inclusion criteria were: aged between 18 and 65 years	20 items/ 4 dimensions	4-point Likert	Treatment management and treatment outcome; Treatment relationship; Communication and information exchange; Reachability of treatment facilities	Administration during an interview	NA	66.1-73.9
Bjertnaes et al, 2015 (123)	To psychometrically test the on-site version of the PIPEQ version (PIPEQ-OS)	Psychiatric Inpatient Patient Experience Questionnaire on-site version (PIPEQ-OS)*	Norway, Norwegian	552 adult psychiatric inpatients	17 items/ 3 dimensions	5-point Likert	Structure and facilities; Patient-centred interaction; Outcomes	Self-completion	NA	74.6
Blais et al, 2002 (50)	To develop a brief unidimensional measure of global satisfaction	Patient Evaluation of Care-5 (PEC-5)*	US, English	109 inpatients from acute psychiatric unit	5 items/ 1 dimension	7-point Likert	Overall patient satisfaction and staff availability and attitude	Self-completion	At discharge	NA
Bramesfeld et al, 2007 (52)	To evaluate the performanc	Tool not named	Germany, German	312 patients recruited in inpatient	NA / 8 dimensions	5- and 4-point Likert +	Dignity; Autonomy; Confidentiality;	Administration during an	NA	NA

	e of mental health care in a catchment area in Germany			and outpatient care facilities Inclusion criteria were: use of "complex mental health services" (i.e. use of social support, medical support or receiving inpatient care) in the catchment area during the past 6 months and be cognitively capable to follow an interview	7 dimensions	numerical responses	Communication; Prompt attention; Quality of basic amenities; Choice; Continuity Access to social support (only for inpatient care)	interview		
Brunero et al, 2009 (51)	To determine the level of consumer satisfaction with an adult acute inpatients mental health	Consumer satisfaction questionnaire*	Australia, English	70 acute psychiatric inpatients in two wards	24 items/ 4 dimensions	10- and 5-point Likert	Quality of care; Staff; Environment and services; Discharge	Self-completion	On the day of discharge	38.5

	service and to cross-sectionally identify key associates of overall satisfaction from within the survey content domain									
Bruyneel et al, 2018 (39)	To describe the development, validation, and findings of a patient experience questionnaire across 7 types of residential and ambulatory mental health care services	Flemish Patient Survey of Mental Healthcare*	Belgium, Dutch	5 168 adult patients from residential or ambulatory services Inclusion criteria were: at least 4 days of admission or at least 4 sessions or contacts Exclusion criteria were: non-Dutch speaking, treated by a self-employed caregiver, cognitively unable to complete the	37 items/ 9 dimensions	11- and 4-point Likert + yes/no format	Information about mental health problems and treatment; Participation; Therapeutic relationship; Personalized care; Organization of care and collaboration between professionals; Safe care; Patient rights; Result and evaluation of care; Discharge management and after-care	Self-completion	NA	NA

				questionnaire						
Caruso et al, 2013 (53)	To develop and perform a primary validation of a questionnaire to assess the main subjective experiences of patients with severe mental illness attending group therapy	Ferrara Group Experiences Scale (FE-GES)*	Italy, Italian	166 patients with severe mental illness attending group therapies in community mental health services Inclusion criteria were: a psychiatric diagnosis according to ICD-10 criteria and to take part in at least 5 group activity sessions of any modality and consent to participate in the study. Exclusion criteria were: diagnosis of mental retardation	20 items/ 5 dimensions	5-point Likert	Sharing of emotions and experiences; Cognitive improvement; Group learning; Difficulty in open expression; Relationships	Self-completion	Prior to discharge	100.0

				(ICD-10 codes F70 to F79).						
Clement et al, 2012 (16)	To develop, and provide an initial validation of, a comprehensive measure for assessing barriers to access to mental health care including a 'treatment stigma' subscale, and to present preliminary evidence about the prevalence of barriers experienced by adults currently or recently using secondary mental health services in the UK	Barriers to Access to Care Evaluation scale (BACE)*	UK, English	117 patients participating in the QUAD study Inclusion criteria were: having received care from secondary mental health services in the 12 months or currently; aged 18 or over; and access to the internet	30 items/ 2 subscales	4-point Likert	Non-stigma related barriers; Stigma-related barriers	Self-completion on-line	NA	NA
Eisen et al, 2001 (54)	To review current	Mental Health Statistics Improvement	US, English	3 443 adults from six	36 items/ 3 dimensions	5-point Likert	Access to care; Quality/appropri	Self-completio	NA	43.1

	trends and national efforts to assess the quality of behavioral health services from the consumer's perspective,	Program's (MHSIP) Consumer Survey*		behavioral health plans (four were public assistance programs and two were commercial plans)			ateness; Outcomes of treatment	n		
Eisen et al, 1999 (55) **Eisen et al, 2001 (54)**	to summarize results of a study to develop a standardized consumer survey for national use, and to discuss implications for assessing the quality of behavioral health and substance abuse services	Consumer Assessment of Behavioral Health Services (CABHS)*	US, English	Inclusion criteria were: to have received at least one behavioral health or substance abuse service in the 12 months prior to the survey	54 items/ 5 dimensions	11-, 4- and 3-point Likert + yes/no format	Getting care quickly; Consumer-provider relationship; Information given by clinicians; Plan access and administrative burden; Waiting more than 15 minutes past appointment time	Self-completion by telephone or mail		42.1
Eisen et al, 2002 (56)	To develop a low-cost, low-burden survey that would address important	Perception of care survey (PoC)*	US, English	6 972 patients treated in inpatient behavioral health or substance	18 items/ 4 dimensions	10-, 4-, 3-point Likert + yes/no format	Information received; Staff-patient relationship; Continuity-coordination of care;	Self-completion	One day before discharge	NA

	quality domains, allow for interprogram comparisons and national benchmarks, be useful for quality improvement purposes, and met accreditation and payer requirements			abuse treatment programs			Global evaluation of care			
Eton et al, 2017 (58) **Eton et al, 2015 (57)**	To develop and validate a new comprehensive patient-reported measure of treatment burden – the Patient Experience with Treatment and Self-Management (PETS)	Patient Experience with Treatment and Self-Management (PETS)*[a]	US, English	332 multi-morbid patients from two clinical sites Inclusion criteria were: ≥21 years old, assigned to a primary care provider at one of the two clinical sites, medical record-confirmed diagnoses of two or more	48 items/ 9 dimensions	5- and 4-point Likert + yes/no format for screening questions	Medical information; Medications; Medical appointments; Monitoring health; Interpersonal challenges; Medical & healthcare expenses; Difficulty with healthcare services; Role and social activity limitations; Physical and mental exhaustion	Self-completion	NA	40.0

				chronic conditions with these diagnoses listed on billing encounters of the past 3 years, and at least one medical record-confirmed encounter with a provider from one of the two clinical sites within the past 18 months for one or more of the selected chronic conditions. Exclusion criteria were: lack of English proficiency.						
Evans et al, 2012 (59)	To develop a patient-reported outcome measure of perceptions of acute	Views on Inpatient Care (VOICE)*	UK, English	360 inpatients from acute wards Inclusion criteria	19 items/ 7 dimensions	6-point Likert	Admission; Care and treatment; Medication; Staff; Therapy and activities;	Self-completion	Relatively soon after admission	45.0

	care			were: participants that could provide informed consent and had been present on the ward for at least 7 days during the 4-week data collection phase			Environment; Diversity			
Forouzan et al, 2014 (60)	To adapt the original form of the Health System Responsiveness Questionnaire, developed by the WHO, to the mental health care system in Iran, by determining the validity and reliability of this new version	Mental Health System Responsiveness Questionnaire (MHSRQ)*	Iran, Farsi	500 mentally ill patients from nine public outpatient clinics Inclusion criteria were: being an adult (18-65 years old), receiving outpatient care during past 12 months, being in remission phase of their disorder and	38 items/ 8 dimensions	5 and 4-point Likert + numerical response s	Attention; Dignity; Clear communication; Autonomy; Effective care; Access to care; Confidentiality; Quality of basic amenities	Administration during an interview	NA	NA

				mentally capable to follow the interview						
Garratt et al, 2006 (61)	To develop and evaluate the Psychiatric Out-Patient Experiences Questionnaire (POPEQ)	Psychiatric Out-Patient Experiences Questionnaire (POPEQ)*	Norway, Norwegian	6 677 psychiatric outpatients attending clinics (aged 18 years and over)	11 items/ 1 dimension and 3 subscales	5-point Likert	Perceived outcome of the treatment; Quality of interaction with the clinician; Quality of information provision	Self-completion postal	NA	43.3-35.2
Olsen et al, 2010 (62)	To apply the Rasch model as a supplement to classical test theory in order to assess the psychometric properties of one of the measures derived from the Psychiatric Out-Patient Experiences Questionnaire (POPEQ)			11 085 psychiatric outpatients attending clinics						
Gensichen et al, 2011 (63)	To evaluate the psychometric properties of the	Patient Assessment of Chronic Illness Care (PACIC)*	US, English	442 patients in primary care Inclusion	20 items/2 dimensions	5-point Likert	Patient activation and problem solving; Goal setting and coordination	Self-completion	At home	91.1

	German PACIC in a sample of patients with major depression	criteria were: (1) diagnosis of major depression with indication for any antidepressive treatment; (2) age 18–80; (3) access to private telephone; (4) ability to give informed consent; (5) ability to communicate in German. Exclusion criteria were: (1) confirmed pregnancy; (2) severe alcohol or illicit drug consumption and (3) acute suicidal ideation assessed by the family

				doctor						
Gigantesco et al, 2003 (64)	To describe the development, the main features and the validation of the ROQ-PW questionnaire	Rome Opinion Questionnaire for Psychiatric Wards (ROQ-PW)*	Italy, Italian	169 inpatients from a psychiatric ward of a general hospital	12 items/ 5 dimensions	5-point Likert	Professional qualities of the staff; Information received; Physical environment; Overall satisfaction; Patient empowerment	Self-completion	6-7 days after admission	97.0
Glick et al, 1991 (65)	This pilot study had the objective of dissecting the process of care in an attempt to understand outcomes for patients with major affective disorders and for their families	Quality Care Intervention Checklist*	US, Japan, Italy	24 patients from three countries (Italy, Japan, US) Inclusion criteria were: 15 to 65 years of age, have had a major affective disorder, either unipolar or bipolar, the acute episode must have been treated between 12 and 18 months before the interview, and have	6 items	Dichotomous	NA	Administration during an interview	NA	NA

				had a family or significant others available at the time of the index episode						
Hansson et al, 1995 (66)	To develop a self-rating patient-satisfaction questionnaire that could be used both as a routine instrument in quality assurance programs and in psychiatric health care services research	Self-rating patient satisfaction questionnaire (SPRI)*	Sweden, Swedish	453 inpatients admitted for more than 3 days 1656 outpatients	43 items/ 7 dimensions 30 items/ 7 dimensions	5-point Likert	Staff-patient relationship; Ward atmosphere and physical milieu; Information; Treatment interventions; Restrictions and compulsory care; Treatment design; Treatment program as a whole Accessibility and availability; Staff-patient relationships; Informational procedures; Patients' influence on treatment planning; Treatment interventions; Treatment design;	Self-completion	Before and after discharge (to be filled in at home within 2 weeks)	46.0-71.0

Hester et al, 2015 (67)	To develop and validate the first brief, service user-centred, English-language instrument (SEQUenCE (SErvice user QUality of CarE)) that be routinely used to assess quality of care in mental health services	SErvice user QUality of CarE (SEQUenCE)*	Ireland, English	61 service users of an independent mental health Inclusion criteria were: aged 18 or older, had a diagnosis of bipolar affective disorder (ICD-10: F31) or a psychotic disorder (ICD-10: F20, 22, 25, 28, 29) from a consultant psychiatrist of at least 6 months standing, and gave informed consent. For inpatients sample, admitted to hospital at least 2 weeks prior their	40 items	5-point Likert	Treatment program NA	Self-completion	Close to discharge	NA

				participation. Exclusion criteria were: current involuntary status, diagnosis of a personality disorder, diagnosis of a primary substance abuse disorder or diagnosis of moderate to severe cognitive impairment						
Howard et al, 2001 (68)	(i) To design instruments to measure satisfaction of consumers, family and support persons and community health providers with mental health services; (ii) To test validity,	Kentucky Consumer Satisfaction Instrument (KY-CSI)*	US, English	189 inpatients from a public psychiatric hospital Inclusion criteria were: older than 18 years, currently hospitalized, and identified by the hospital treatment	19 items/ 1 dimension and 3 subscales	5 and 4-point Likert	Environment; Affiliation or esteem; Growth or self-actualization	Administration during an interview	Completion near the patient's discharge (generally within 24 to 72 hours before leaving)	100.0

	reliability and feasibility of the instruments before statewide implementation; and (iii) To use study findings for recommendations about quality improvement in public psychiatric hospitals			team as symptomatically stabilized. Exclusion criteria were: identified by the hospital treatment team as incapable of providing informed consent, physical condition that precluded participation, or forensic classification						
Ivarsson et al, 2007 (69) **Ahlfors et al, 2001 (70)**	To validate a patient self-rating version of the UKU consumer satisfaction scale (Pat-UKU-ConSat) in relation to the original interviewer version, and to analyze its internal	A patient self-rating version of the UKU-Consumer Satisfaction scale (Pat-UKU-ConSat)*	Sweden, Swedish	37 patients from five psychiatric services	12 items/ 1 dimension and 2 subscales	7-point Likert + yes/no format for screening questions and VAS	Structure and process; Outcome	Self-completion	NA	NA

	consistency									
Jenkinson et al, 2002 (71)	To design a core set of items from the Picker adult in-patient questionnaire, a short form of the original, which could be used to make comparisons between hospitals and for monitoring trends over time	Picker Patient Experience questionnaire (PPE-15)*[a]	UK, English.	Patients who had attended acute care hospitals in five countries: 2249 in UK, 2663 in Germany, 3274 in Sweden, 7163 in Switzerland and 47576 in USA	15 items/ 7 dimensions	6-, 5-, 4- and 3-point Likert + yes/no format for screening questions	Information and education; Coordination of care; Physical comfort; Emotional support; Respect for patient preferences; Involvement of family and friends; Continuity and transition	Self-completion postal	Completion within 1 month of discharge	46.0-74.0
Joyce et al, 2010 (72)	To examine the psychometric characteristics of the final draft 43-item measure of perceived COC—the Alberta Continuity of Services Scale-Mental Health (ACSS-	Alberta Continuity of Services Scale-Mental Health (ACSS-MH)*	Canada, English	441 patients with a severe and persistent mental illness who were receiving mental health services and who participated in a 18-month longitudinal follow-up study	32 items/ 3 dimensions	5-point Likert	Responsive system; Individualized care; Responsive caregiver	Self-completion	NA	NA

	MH)			Inclusion criteria were: a confirmed diagnosis of severe mental illness (psychotic disorder, bipolar disorder, or unipolar depressive disorder of at least 24 months duration) and age between 18 and 65 years. Exclusion criteria were: being under guardianship or receiving involuntary or forensic care.						
Kertesz et al, 2014 (73)	To portray the process and psychometrics supporting a new	Primary Care Quality-Homeless (PCQ-H)*	US, English	563 homeless-experienced clients across 3 VA facilities and 1 non-VA	33 items/ 1 dimension and 4 subscales	4-point Likert	Patient clinician relationship; Cooperation; Access & coordination; Homeless-specific needs	Self-completion	NA	NA

	survey tool focused on primary care for homeless individuals; and to provide a portrait of the combined qualitative and quantitative procedures that can support the development of patient-reported care surveys for patient populations with unique concerns and needs			Health Care for the Homeless Program Inclusion criteria were: evidence of past or current homelessness and 2 or more visits to a primary care provider in the past 2 years.						
Kolb et al, 2000 (74)	To investigate the psychometric properties of a survey designed specifically to evaluate consumer satisfaction with	Inpatient Psychiatric Questionnaire*	US, English	1 351 inpatients from psychiatric units	34 items/ 6 dimensions	NA	Nonclinical services; Psychiatric Care; Staff; Medical outcome; Patient education; Program components/activities	Self-completion (with the option of returning the questionnaire immediately or sent by mail)	On the day of discharge	53.0

	inpatient behavioral health care across diverse facilities									
Larsen et al, 1979 (75)	To develop and shape a general scale to assess client/patient satisfaction	Client Satisfaction Questionnaire (CSQ-8)*[a]	US, English	248 psychiatric outpatients	8 items/ 1 dimension	4-point Likert	NA	Self-completion	NA	NA
Lelliott et al, 2001 (76)	To develop and test a self-assessment	Carers' and User's Expectations of Services - User Version (CUES-U)*	UK, English	449 users of local mental health services	16 items/ 3 dimensions	3-point Likert	Quality of interactions with mental health workers;	Self-completion	NA	NA
Blenkiron et al, 2003 (77)	instrument to enable users of mental health services to rate their experience across the range of domains that they consider to be important			86 adult working age patients. Inclusion criteria were: aged between 16 and 65 years, had a mental health disorder of greater than 6 months' duration and receiving input from at least one mental			Sense of alienation; Finance, daytime activities and social relationships			72.0

				health professional as part of the care program approach						
Llyod-Evans et al, 2010 (78)	To enhance understanding of how to measure content of care by developing and evaluating four instruments, each using a different measurement method	Camden Content of Care Questionnaire – Patient version (CCCQ-P)*	UK, English	314 patients from four alternative residential acute services and four standard acute inpatient wards (three non-hospital crisis houses and five inpatient wards)	21 items/ 4 dimensions	7-point Likert + yes/no format	Social intervention; Psychological intervention; Physical and pharmacological intervention; General care and organization	Self-completion	At the time of the patient's discharge	70.2
MacInnes et al, 2010 (79)	To develop and validate a scale designed to measure satisfaction with forensic mental health services	Forensic Satisfaction Scale (FSS)	UK, English	63 inpatients in forensic medium and low secure units	60 items/ 7 dimensions	5-point Likert	Staff interaction; Rehabilitation; Milieu; Communication; Finance; Safety; Overall satisfaction	Self-completion	NA	53.8
Madan et al 2014 (80)	To develop the Menninger Quality of Care (MQOC)	Menninger Quality of Care (MQOC)*	US, English	337 adult inpatients	20 items/ 4 dimensions	4-point Likert	Alliance with treatment providers; Family engagement; Discharge	Self-completion	24 to 48 hours before discharge	80.8

	measure						planning; Treatment enhanced self-efficacy			
Mavaddat et al, 2009 (81)	To develop a single valid and reliable questionnaire relevant to people with SMI or common mental health problems that could assess patient experience of primary care mental health at the practice level for patients with mental health problems	Patient Experience Questionnaire (PEQ)*	UK, English	241 patients seen in one of the nine study practices during the past 3 months with a diagnosis of depression, bipolar disorder, obsessive-compulsive disorder, schizophrenia, anxiety or stress Exclusion criteria were: not suitable to complete a questionnaire due to an acute illness or recent bereavement for example	20 items/ 2 dimensions	5-point Likert	Attributes of the GP; Experiences with the practice in general	Self-completion	NA	25.1
Mayston et al, 2017 (82)	To explore the dimensions and	Mental Health Service Satisfaction scale (MHSSS)*	Ethiopia, English and Amharic	200 and 150 service users with severe mental	24 items/ 1 dimension	4-point Likert	NA	Administration during an interview	NA	NA

	meaning of satisfaction with services and develop a testable measure among users of a psychiatric nurse delivered out-patients service for people living with severe mental disorders			disorder						
McGuire et al, 2007 (83)	To develop a measure of the therapeutic relationship (TR) with clinician and patient versions using psychometric principles for test construction	Scale To Assess the Therapeutic Relationship - Patient version (STAR-P)*	UK, English	266 patients Inclusion criteria were: aged 18–65, in the care of a community mental health team and had severe mental illness	12 items/ 3 dimensions	5-point Likert	Positive clinician input; Non-supportive clinician input; Positive collaboration	Self-completion	NA	NA
Meehan et al, 2002 (84)	To describe the development and testing of a brief satisfaction	Inpatient Evaluation of Service Questionnaire (IESQ)*	Australia, English	356 inpatients from acute psychiatric units and rehabilitatio	20 items/ 3 dimensions	5-point Likert	Staff-patient alliance; Satisfaction with environment; Satisfaction with treatment	Self-completion	Who were approaching discharge	72.0

	measure for inpatients, the Inpatient Evaluation of Service Questionnaire			n facilities Exclusion criteria were: if stay less than 7 days and readmitted patients						
Misdrahi et al, 2009 (85)	To build a self-rating scale easy to use in clinical psychiatric practice to assess therapeutic relationship (TR), including drug-taking aspects and the relationship with the clinician on a day-to-day basis. A secondary objective was to assess the scale's validity and the association between TR and	4-Point ordinal Alliance Scale (4PAS)*	France, French	84 inpatients in the acute unit of three psychiatric hospitals Inclusion criteria were: older than 18 years, clinically diagnosed with schizophrenia or schizoaffective disorders, sufficiently capable of understanding the protocol; and give informed consent	11 items/ 2 dimensions	4-point Likert	Empathy experienced; Psychoeducation	Self-completion	Completion 1 week before discharge, after the remission of acute symptomatology	91.3

	adherence									
Moutoussis et al, 2000 (86)	(i) To assess the quality of psychiatric outpatient care using the Psychiatric Care Satisfaction Questionnaire; (ii) To identify patient characteristics that predict overall satisfaction	Psychiatric Outpatient Satisfaction Questionnaire*	UK, English	82 outpatients from two clinic sites (five general adult and three old age psychiatry teams)	34 items/ 4 dimensions	5-point Likert	General statements; Empowerment; Choice and treatment; Quality of doctor-patient relationship	Self-completion in the department or at home	NA	52.9
Nabati et al, 1998 (87)	To describe the psychometric properties of a simple patient satisfaction self-report instrument originally developed for use in primary care patients, adapted for use in mental health clinic	Satisfaction Index - Mental Health (SI-MH)*	US, English	144 veterans of various psychiatric diagnoses attending mental health clinics Inclusion criteria were: at least their second visit to their mental health provider	12 items/ 1 dimension	6-point Likert	NA	Self-completion	NA	88.0

	patients of varied educational and socioeconomic backgrounds										
Nordon et al, 2014 (88)	To describe the development of the PASAP scale in French and its psychometric properties	Patient Satisfaction with Psychotropic (PASAP)*	France, French	314 bipolar outpatients from the French subgroup of an observational study cohort "European Mania in Bipolar Longitudinal Evaluation of Medication" (EMBLEM) Inclusion criteria were: age >18, acute manic or mixed episode, a need for initiation or switch of psychotropic medication and patient	9 items/ 1 dimension	5-point Likert	NA		Self-completion at patient home	At 3 months after psychotropic drug change	68.4

				consent						
Oades et al, 2011 (42)	To develop a consumer satisfaction questionnaire in which consumers work as collaborative researchers to increase its face validity and relevance	Consumer Evaluation of Mental Health Services (CEO-MHS)*	Australia, English	202 mental health consumers from public and non-government mental health service centres	26 items/ 2 dimensions	5-point Likert	Empowerment; Dehumanization	Self-completion	NA	NA
Rose et al, 2011 (89)	To build on the consumer-centered survey development of Oades and to develop further its applicability to NGO mental health services in Australia, while keeping the focus on the consumer involvement	A short form of Consumer Evaluation of Mental Health Services (CEO-MHS)*	Australia, English	481 mental health service consumers of two non-government mental health and disability services	13 items/ 1 dimension	5-point Likert	NA.	Administration during an interview		Greater than 50%
Ortiz et al,	(i) To	Inpatient Consumer	US, English	34 878	28 items/ 6	5-point	Outcome;	Self-	At	51.0

2012 (90)	further re-evaluate, through EFA and CFA, the structure of an instrument intended to measure consumers' satisfaction with care in psychiatric settings and (ii) To examine and publish the psychometric characteristics, validity and reliability, of the Inpatient Consumer Survey	Survey (ICS)*		individuals (aged 13 years and older) from psychiatric hospitals whose survey data were extracted from the Behavioral Healthcare Performance Measurement System (BHPMS) Exclusion criteria were: if hospitals were not enrolled in BHPMS at the time of the study	dimensions	Likert	Dignity; Rights; Participation; Environment; Empowerment	completion	discharge and at annual review	
Parker et al, 1996 (91)	To describe the first stage of development of patient satisfaction form designed	Patient satisfaction Survey*	Australia, English	172 psychiatric outpatients	43 items/ 4 dimensions	4-point Likert	Doctor-patient alliance; Patient's privacy; Amenities; Service	Self-completion	NA	NA

	for psychiatric outpatients									
Pellegrin et al, 2001 (92)	To evaluate the reliability and preliminary validity of the Charleston Psychiatric Outpatient Satisfaction Scale	Charleston Psychiatric Outpatient Satisfaction Scale (CPOSS)*	US, English	282 adult outpatients treated in clinics affiliated with a public-academic psychiatric institution	15 items	5- and 4-point Likert	NA	Self-completion	NA	NA
Perreault et al, 2001 (93)	To verify whether information on services would appear as a distinct dimension of satisfaction in a multidimensional scale	Opinion Questionnaire on Outpatient Services (OQOS-21)*	Canada, French	Adult patients from two outpatient clinics (n=263 for the original version and n=200 for the adapted version)	21 items/ 5 dimensions	4-point Likert	Therapeutic alliance; Respect towards patient; Service reception and access; Information on services and treatment; Clinic location and atmosphere	Administration during an interview	NA	74.5
Perreault et al, 2006 (94)	To develop a scale in order to determine the informational needs deemed most important by	Patients' Perspective on Information Questionnaire (PPIQ)*	Canada, French	86 adult psychiatric patients in two outpatient clinics Inclusion criteria were: patients	18 items/ 2 dimensions	3-point Likert	Information; Satisfaction	Administration during an interview	NA	68.8

	psychiatric outpatients and to determine their level of satisfaction with information received			were approached if they had at least 4 consultations with their primary therapist, if they were not in a state of crisis and after provided a written or verbal consent						
Peytremann-Bridevaux et al, 2006 (95)	To compare two psychiatric-specific and one generic questionnaires assessing patients' satisfaction after a hospitalization in a psychiatric hospital	Saphora-Psy (developed by Comité de Coordination de l'Evaluation Clinique et de la Qualité en Aquitaine – CCECQA)	France, French	Adult psychiatric inpatients	34 items/ 5 dimensions	5 and 3-point Likert + yes/no format for screening questions	Continuity of care; Nursing care; Medical care; Relation with other patients; Services and performance	Self-completion	Before discharge	NA
Phattharayuttawat et al, 2005 (96)	To develop the Thai Psychiatric Satisfaction Scale (TPSS), having adequate and	Thai Psychiatric Satisfaction Scale (TPSS)	Thailand, Thai	384 psychiatric outpatients Inclusion criteria were: age between 15-60 years,	84 items/ 7 dimensions	5-point Likert	Professionals' Skills and Behavior; Information; Access; Efficacy; Types of Intervention; Relative's	Self-completion	NA	NA

	sufficient validity and reliability			able to communicate, contact with mental health services during the 3-month period preceding the start of the study. Exclusion criteria were: a disturbed condition, aggressive behavior, suicidal attemps, mental retardation, and primary dementia or other severe organic disorders			Involvement; Environment and Setting			
Priebe et al, 1993 (97)	To assess the quality of the helping alliance between patients and clinical case managers in psychiatric community	Helping Alliance Scale (HAS)*	Germany, German	72 patients receiving long-term treatment in psychiatric community care	5 items	100-mm long VAS ranging from 0 (“not at all”) to 100 (“entirely") and one	NA	Administration during an interview	NA	72.0

	care and to examine its value as a predictor of treatment outcome					categoric al response				
Rofail et al, 2005 (98)	To produce a reliable and valid standardize d measure of patients' satisfaction with antipsychoti c medication	Satisfaction with antipsychotic medication scale (SWAM)*	UK, English	315 patients on the caseload of local mental health services Inclusion criteria were: case-note diagnosis of schizophreni a; currently taking antipsychoti c medication or advised by doctor to do so; aged between 16-65 years and written consent. Exclusion criteria were: case-note diagnosis of comorbid learning disabilities and brain	23 items/ 2 dimensions	5-point Likert	Treatment acceptability; Medication insight	Self-completio n	NA	40.0

				injury or any other organic disorder.						
Rose et al, 2009 (99)	(i) To develop a user-generated measure of continuity of care, (ii) To establish its test-retest reliability and (iii) To test it in a field trial sample	CONTINUity of care - User Measure (CONTINU-UM)	UK, English	167 service users recruited from community mental health teams Inclusion criteria were: a diagnosis of psychosis, in touch with services for at least 2 years and aged 18-65 years	32 items/ 2 dimensions and 16 subscales	5-point Likert	Accessing services; Range of services; Waiting; Out of hours support; Hospital discharge; Staff changes; Information; Flexible level of support; Individual progress; Day centers; Care plans; Crisis; Communication between staff; Support from other users; Repeating your life history; Avoiding services	Self-completion	NA	NA
Røssberg et al, 2003 (100) **Røssberg et al, 2003 (101)**	(i) To examine the psychometric properties of the two subscales of the Ward Atmosphere Scale	Ward Atmosphere Scale – Real ward (WAS-R)	Norway, Norwegian	550 inpatients on 54 wards for psychotic patients	80 items/ 11 subscales	4-point Likert	Involvement; Support; Spontaneous behavior; Autonomy; Practical orientation; Personal problem	Self-completion	NA	NA

	(WAS). (ii) To re-evaluate the psychometric properties of the other eight subscales of the Ward Atmosphere Scale						orientation; Angry and aggressive behavior; Order and organization; Program clarity; Staff control; Staff attitude to expressed feelings			
Ruggeri et al, 2000 (102) **Ruggeri et al, 1993 (124)**	To describe the development, translation, cultural validation and reliability of a new European Version of the VSSS (VSSS-EU), for use in multi-site international comparative studies	Verona Service Satisfaction Scale - European version (VSSS-EU)*	Italy, Italian	399 adult patients across five EPSILON sites (the Netherlands, Denmark, England, Spain and Italy) Inclusion criteria were: aged 18-65 years, an ICD-10 diagnosis of F20, in contact with mental health services during the 3-month period preceding the start of the study.	54 items/ 7 dimensions	5- and 3-point Likert + yes/no format for screening questions	Overall satisfaction; Professionals' skills and behavior; Information; Access; Efficacy; Types of intervention; Relatives' involvement	Self-completion	NA	NA

				Exclusion criteria were: current residence in prison, secure residential services or hostels for long-term patients, co-existing learning disability, primary dementia or other severe organic disorder and extended inpatient treatment episodes than 1 year						
Rush et al, 2013 (103)	To test important psychometric properties in terms of validity and reliability in order to evaluate the appropriateness of the tool for assessing various	Ontario Perception of Care Tool for Mental Health and Addictions (OPOC-MHA)*	Canada, English	1 476 clients (aged 12 years or older) with mental health or/and addiction problems receiving services in one or more of the programs	39 items/ 4 dimensions	4-point Likert	Recovery; Services; Access/entry to services; Discharge/ leaving the program	Self-completion	The timing of tool administration varied by agency with some agencies administering the tool at program completion and others conducting a one-day	NA

	aspects of client perceptions of care in mental health and addiction treatment services and the utility for quality improvement			being offered of the 22 pilot sites. Exclusion criteria were: immediate need for crisis services					or one-month blitz	
Schalast et al, 2008 (104)	To validate a short questionnaire, designed for assessing forensic psychiatric wards	Essen Climate Evaluation Schema (EssenCES)*	Germany, German	327 patients from 17 forensic mental hospitals	15 items/ 3 dimensions	5-point Likert	Therapeutic hold; Experienced safety; Patients' cohesion and mutual support	Self-completion	NA	NA
Schröder et al, 2007 (105) **Schröder et al, 2010 (106)**	To test the psychometric properties and dimensionality of a new instrument, quality in psychiatric care (QPC), and to describe and compare quality of care among	Quality in Psychiatric Care –Inpatient (QPC–IP)	Sweden, Swedish	265 inpatients from eight general adult psychiatric wards Inclusion criteria were: able to understand and express themselves in Swedish, and had at least a three-	30 items/ 6 dimensions	4-point Likert	Encounter; Participation; Discharge; Support; Secluded environment; Secure environment	Self-completion	On the day of discharge	NA

	inpatients as measured by this instrument			day stay in the ward before discharge						
Schröder et al, 2007 (105) **Schröder et al, 2011 (107)**	To test the psychometric properties and dimensionality of the Quality in Psychiatric Care – Outpatient (QPC–OP) instrument	Quality in Psychiatric Care – Outpatient (QPC–OP)	Sweden, Swedish	1340 adult outpatients from general psychiatric clinics	30 items/ 8 dimensions	4-point Likert	Encounter; Participation-empowerment; Participation-Information; Discharge; Support; Environment; Next of kin; Accessibility	Self-completion	Before leaving the clinic	61.0
Shiva et al, 2009 (108)	To develop and investigate the psychometric properties of a survey designed to evaluate patient satisfaction on forensic and civil psychiatric inpatient units	Inpatient Satisfaction Questionnaire (ISQ)*	US, English	843 male adult inpatients from forensic and civil psychiatric units (hospitalized for at least 5 days). Non-inclusion of floridly psychotic or recently violent patients	23 items 14 items for F-ISQ scale/ 4 dimensions 11 items for C-ISQ scale/ 2 dimensions	5-point Likert	Medication and treatment; Physical environment; Telephone access; Unit rules and procedures Needs and Opportunities; Food and Comfort	Administration during an interview	NA	100.0
Slade et al, 2014 (109)	To develop and evaluate a brief quantitative	Clinical Decision-making Involvement and Satisfaction - version Patient (CDIS-P)*	UK, English	443 service users Inclusion criteria	7 items/ 2 subscales	5- and 3-point Likert	Satisfaction; Involvement	Self-completion	NA	NA

	five-language measure of involvement and satisfaction in clinical decision-making (CDIS) – with versions for patients (CDIS-P) and staff (CDIS-S) - for use in mental health services	were: aged 18-60, sufficient command of the local language, have a primary research diagnosis of mental disorder, cognitive ability to give informed consent and complete study measures, expected contact with services during the study period and presence of a severe mental illness for at least two years Exclusion criteria: diagnosis of learning disability, dementia, substance

				abuse or organic brain disorder						
Slater et al, 1982 (110)	To describe a Satisfaction with Mental Health Care (SMHC) scale	Satisfaction with Mental Health Care (SMHC)*	US, English	170 patients from an outpatient mental health clinic	32 items/ 4 dimensions	4-point Likert	Overall care; Therapeutic relationship; Prevention; Planning access	Self-completion	NA	NA
Speckens et al, 2000 (111)	To develop a questionnaire that assesses the reassurability of patients and to examine its psychometric qualities	Reassurance Questionnaire (RQ)*[a]	The Netherlands, Dutch	204 subjects from the general population, 113 general practice patients, 130 general medical outpatients and 183 general medical patients with unexplained physical symptoms	8 items/ 1 dimension	4-point Likert	NA	Self-completion postal	NA	74.0-97.0
Svedberg et al, 2008 (113) **Svedberg et al, 2007 (112)**	To develop and investigate psychometric properties in terms of factor structure, internal consistency, and test–	Health Promotion Intervention Questionnaire (HPIQ)*	Sweden, Swedish	135 outpatients in contact with eight mental health services Inclusion criteria were:	19 items/ 4 dimensions	5-point Likert	Alliance; Empowerment; Educational support; Practical support	Administration during an interview.	NA	56.5

	retest reliability of the Health Promotion Intervention Questionnaire (HPIQ), a newly developed instrument intended to measure patients' subjective experiences of a health promotion intervention in the mental health services			experience of outpatient care, understanding of and ability to read the Swedish language, and more than 18 years of age.						
Uijen et al, 2011 (114) **Uijen et al, 2012 (115)**	To develop and pilot test a generic questionnaire to measure continuity of care from the patient's perspective across primary and secondary care settings	Nijmegen Continuity Questionnaire (NCQ)*[a]	The Netherlands, Dutch	288 and 268 patients with one or more chronic disease recruited from general practice and hospital/out patient department Exclusion criteria were: under the age of	28 items/ 3 subscales	5-point Likert	Care provider knows me; Care provider shows commitment; Team/cross-boundary continuity	Self-completion at home	NA	72.0-76.0

				18 years or who were unable to speak or read Dutch						
Ul-Haq, 2012 (116)	To evaluate patients' satisfaction with a psychiatric day hospital in the West Galway Catchments area	Service Satisfaction Questionnaire*	UK, English	37 patients (18 years and over) who attended the day hospital	33 items	5-point Likert	NA	Self-completion in the day hospital or at home	NA	84.1
Ware et al, 2003 (117)	To support and strengthen research on continuity of care in mental health services through the development of a formal measure	CONNECT*	US, English	400 adults with serious mental illness who were actively using public mental health services Inclusion criteria were: a diagnosis of serious mental illness; aged of 18 or greater and current relationships with at least two mental	59 items/ 5 dimensions and 13 subscales + one single-item indicator	5-point Likert	Knowledge; Flexibility; Availability; Coordination; Transitions	Administration during an interview	NA	NA

				health practitioners						
Webster et al, 2012 (118)	To identify aspects of mental health nursing care that are most likely to influence satisfaction with patients who are experiencing anxiety in two private mental health care setting	Patients Satisfaction Questionnaire*	Australia, English	189 inpatients who had experienced anxiety upon admission to two private mental health settings	15 items/ 7 dimensions	5-point Likert	Information and education; Anxiety intervention; Environmental comfort; Emotional support; Respect for patient preferences; Involvement of family, friends; Continuity of care at discharge	Self-completion	At discharge	94.0
Wongpakaran et al, 2013 (119)	To examine the psychometric properties of the seven-item Group Cohesiveness Scale	Group Cohesiveness Scale (GCS)*	Thailand, Thai	96 psychiatric inpatients attending a group therapy session and whose condition had stabilized Exclusion of patients who are likely to disturb group therapy sessions	7 items/ 1 dimension and 2 subscales	5-point Likert	Cohesiveness; Engaged	Self-completion	To complete at the end of the group therapy session	NA

Woodring et al, 2004 (120)	(i) To quantify patient perceptions of staff attributes, the therapeutic environment, and treatment gains during their hospitalization, (ii) To establish acceptable reliability and validity for this population-specific measure, and (iii) To identify the most and least helpful aspects of care	Penn State Inpatient Psychiatry Satisfaction Survey (PSIPSS)*	US, English	673 inpatients in a psychiatric closed unit	15 items/ 2 dimensions	5-point Likert	Professional care and milieu; Treatment issues	Self-completion	On the day of discharge	70.0
Wright et al, 2006 (125)	To compare patient satisfaction of male and female users of Veterans Health Administration (VHA)	Survey of Health care Experiences of Patients (SHEP)*	US, English	107 995 outpatients and 112 817 inpatients from Veterans' Health Administration services	I-SHEP: 42 items/ 9 dimensions	yes/no or multiple-choice responses	Access; Courtesy; Education and information; Coordination of care; Attention to patient preferences; Emotional	Self-completion by mail	Who were discharged	94.1

	services						support; Family involvement; Physical comfort; Preparation for transition to outpatient care			
					O-SHEP: 43 items/ 11 dimensions		Access; Continuity; Courtesy; Education and information; Overall coordination; Visit coordination; Specialist care; Pharmacy pick-up; Pharmacy mail; Attention to patient preferences; Emotional support			98.4
Zendjidjian et al, 2015 (121)	To develop a specific French self-administered instrument for measuring hospitalized patients' satisfaction in	Satisfaction with Psychiatry Care Questionnaire-22 (SATISPSY-22)*	France, French	270 psychiatric inpatients from two psychiatric hospitals	22 items/ 6 dimensions	5-point Likert	Staff; Quality of care; Personal experience; Information; Activity; Food	Self-completion	Before leaving the hospital	91.5

	psychiatry based on exclusive patient point of view, according to the psychometric standards									
Zimmerman et al, 2017 (122)	To develop a scale that would be psychometrically sound, provide information that clinicians could use to modify their behavior, be sensitive enough to distinguish amongst clinicians, and be brief enough so that patients would not perceive scale completion as overly burdensome	Clinically Useful Patient Satisfaction Scale (CUPSS)*	US, English	412 psychiatric outpatients 500 partial hospital patients	18 items/ 4 dimensions 16 items/ 4 dimensions	5-point Likert	Clinician attitude and behavior; Office environment and staff; Global satisfaction; Expectation of improvement	Self-completion	At the end of the initial visit	76.3-86.8

* Items are available in the article.

[a] Generic instruments.

Abbreviation: NA, not available.

Table 2. Development and performance characteristics of the instruments.

Reference	PREM developed	Viewpoint	Questionnaire development	Internal consistency (range)[a]	Construct validity
Aloba et al, 2014 (43) **Anderson et, 1990 (44)**	Trust in Physician Scale (TPS)*	Review of literature, Professionals' and patients' viewpoints	Review of the literature, Interviews with patients and various health care providers, Item selection based on item analysis and meaning	0.68 (0.66-0.76)	2 Factors (48%) Corrected item-to-scale correlations ranged from 0.30 to 0.52
Atkinson et al, 2004 (45)	Treatment Satisfaction Questionnaire for Medication (TSQM)*	Review of literature, Patients' viewpoint	A thorough review of the scientific literature, Three focus group sessions with patients (n=30), Item selection based on in-depth patient interviews (n=17) and factor analyses	(0.86-0.90)	4 factors (three first-order factors (75.6%) and one second-order factor (79.1%)) Inter-scale correlations ranged from 0.35 to 0.72
Baker, 1990 (46)	Consultation Satisfaction Questionnaire (CSQ)*	Review of literature, Professionals' and patients' viewpoints	An initial review of other questionnaires on patient satisfaction together with general practice studies that included surveys of patient opinions, Discussion with fellow general practitioners, Patients' comments of their personal experience on their care, Item selection based on item analysis and a factor analysis	0.91 (0.67-0.87)	3 factors Correlations between each factor and the general satisfaction subscale ranged from 0.50 to 0.64
Barker et al, 1999 (47) **Barker et al, 1996 (48)**	Psychiatric Care Satisfaction Questionnaire (PCSQ)*	Review of literature, Professionals' and patients' viewpoints	Some items were adapted from several satisfaction questionnaires and previous research findings, Clinical impressions of psychiatric and nursing staff, Patient interviews during the initial pilot work (involving 15 or 20 day hospital patients)	0.82-0.89	NA
Berghofer et al, 2011 (49)	Evaluation of Client Services (ECS)*	Review of literature,	Translation of the items from the German WPI ("Vienna Patient	0.93 (0.79-.0.92)	4 factors (66%) Corrected item-total

		Experts' viewpoint	Satisfaction Inventory") into English, Wording improvements and addition of three new items due to their relevance for US-specific treatment aspects based on the input of experts in mental health, epidemiologists and questionnaire development experts, Item selection based on factor analysis		correlations ranged from 0.37 to 0.79. Inter-correlations of ECS subscales ranged from 0.19 to 0.76
Bjertnaes et al, 2015 (123)	Psychiatric Inpatient Patient Experience Questionnaire on-site version (PIPEQ-OS)*	Review of literature	The PIPEQ-OS is a revised version of an existing questionnaire originally developed for post-discharge measurement which was updated to reflect the latest developments of the national program	0.79-0.91	3 factors Item-total correlations ranged from 0.46 to 0.83
Blais et al, 2002 (50)	Patient Evaluation of Care-5 (PEC-5)*	Review of literature, Professionals' and patients' viewpoints	Review of the literature, Focus group discussions involving both unit staff and patients, Item selection based on rational review, empirical analyses and Rasch refinement	0.88	1 factor (70%) Item to scale correlations ranged from 0.61 to 0.79
Bramesfeld et al, 2007 (52)	Tool not named	Review of literature, Patients' viewpoint	Items were derived from the German version of the Multi-Country Services Survey (MCSS) questionnaire which was tailored to suit mental health care, Previous qualitative work with service users	NA	NA
Brunero et al, 2009 (51)	Consumer satisfaction questionnaire*	Review of the literature, Professionals' and patients' viewpoints	A series of discussions groups with consumers and health staff, Literature review on patient satisfaction and national mental health standards	NA	NA
Bruyneel et al, 2018 (39)	Flemish Patient Survey of Mental Healthcare*	Review of literature, Professionals' and	A scoping literature review, Focus group-type brainstorming sessions with patient representatives	NA	9 factors

		patients' viewpoints	(n=6) and health care professionals (n=18)		
Caruso et al, 2013 (53)	Ferrara Group Experiences Scale (FE-GES)*	Professionals' viewpoint	Focus groups were held four times in three community mental health services with clinicians from different professional backgrounds, Item selection based on factor analyses	0.85 (0.70-0.85)	5 factors (60.8%) Inter-item correlations <0.80 and inter-subscale correlations ranged from 0.059 to 0.271
Clement et al, 2012 (16)	Barriers to Access to Care Evaluation scale (BACE)*	Review of literature, Patients' and professionals' viewpoints	An ongoing systematic review of the literature on stigma and healthcare seeking, Items added by the research team based on their knowledge of the research literature and from the expert panel's suggestion, Item selection based on comments of an expert panel, psychometric analyses and further discussion within the research team	0.89 for the Treatment stigma subscale	Inter-item correlations did not exceed >0.7 (data not shown)
Eisen et al, 2001 (54)	Mental Health Statistics Improvement Program's (MHSIP) Consumer Survey*	Review of literature, viewpoint of a wide range of stakeholders	Review of published and unpublished consumer surveys, MHSIP survey was developed by a task force assembled and supported by the Center for Mental Health Services	0.81-0.91	3 factors (47%) Corrected item-total correlations ranged from 0.39 to 0.73
Eisen et al, 2001 (54) **Eisen et al, 1999 (55)**	Consumer Assessment of Behavioral Health Services (CABHS)*	Review of literature, viewpoint of a wide range of stakeholders	CAHPS survey was used as foundation, An extensive literature review including published and unpublished surveys, Input from focus groups, Refinement of a questionnaire thought review by experts and consumer focus groups (n=3), cognitive interviews and a pilot test	0.55-0.87	5 factors (60%) Corrected item-total correlations ranged from 0.34 to 0.74
Eisen et al, 2002 (56)	Perception of care survey (PoC)*	Review of literature	Review of existing measures of consumer assessment of inpatient and outpatient medical and psychiatric care	(0.58-0.83)	4 factors (72%)

Eton et al, 2017 (58) **Eton et al, 2015 (57)**	Patient Experience with Treatment and Self-Management (PETS)*	Patients' viewpoint	A conceptual framework was drawn from a previous work based on qualitative interviews (n=50) and four focus groups with patients (n=25), Item selection based on feedback from a stakeholder panel, cognitive pre-testing with patients (n=23) and psychometric analyses	0.79-0.95	9 factors
Evans et al, 2012 (59)	Views On Inpatient Care (VOICE)*	Patients' viewpoint	Repeated focus groups of service users, Item selection based on relevance and item analysis	0.92	NA
Forouzan et al, 2014 (60)	Mental Health System Responsiveness Questionnaire (MHSRQ)*	Review of literature, Patients' and professionals' viewpoints	Translation of the original English version of the Health System Responsiveness Questionnaire into Farsi, Modifications according to the findings of previous qualitative studies with providers and service users	(0.56-0.92)	8 factors Item-rest correlations ranged from 0.21 to 0.86 Inter-scale correlations ranged from 0.14 to 0.73
Garratt et al, 2006 (61) **Olsen et al, 2010 (62)**	Psychiatric Out-Patient Experiences Questionnaire (POPEQ)*	Review of literature, Patients' viewpoint	A literature review, Patient interviews (n=12) and piloting	0.91-0.92 (0.81-0.87)	1 factor (53.7-57.3%) Item-total correlations ranged from 0.51 to 0.79
Gensichen et al, 2011 (63)	Patient Assessment of Chronic Illness Care (PACIC)*	Review of literature, Professionals' viewpoint	Items selected from a larger pool of items generated by a national pool of experts and the Chronic Care Model framework, Item selection	0.91(0.45-0.80)	2 factors (46.5%) Corrected item-scale correlations ranged from 0.24 to 0.69
Gigantesco et al, 2003 (64)	Rome Opinion Questionnaire for Psychiatric Wards (ROQ-PW)*	Review of literature, Patients' viewpoint	Review of the literature, An open-questionnaire survey conducted among psychiatric inpatients (n=40), Item selection based on psychiatrists and focus groups' opinions	0.82 (0.35-0.71)	3 factors (67.2%)
Glick et al, 1991 (65)	A Quality Care Intervention Checklist*	NA	NA	NA	NA
Hansson et al, 1995	Self-rating patient satisfaction	Review of	An ad hoc questionnaire was used as	0.84-0.87	Correlation inter-

(66)	questionnaire (SPRI)*	literature, Patients', relatives' and professionals' viewpoints	a starting point, A series of hearings with patients, patient organizations, relatives' organizations, and professionals from the field of psychiatry, Item selection based on findings from pilot studies		subscales ranged from 0.04 to 0.22
Hester et al, 2015 (67)	SEQUenCE (SErvice user QUality of CarE)*	Review of literature, Patients' viewpoint	Six focus groups with patients (n=29), A review of the clinical practice guidelines, Item selection based on a field-testing and psychometric analyses	0.87	Inter-item correlations <0.80 (data not shown)
Howard et al, 2001 (68)	Kentucky Consumer Satisfaction Instrument (KY-CSI)*	Review of literature, Professionals' and patients' viewpoints	Maslow's theory was used as a conceptual framework, Items were derived from the literature findings, previously collected data from focus groups with hospitalized consumers, Consumer and clinician input	0.90 (0.69-0.82)	1 factor
Ivarsson et al, 2007 (69) **Ahlfors et al, 2001 (70)**	A patient self-rating version of the UKU-Consumer Satisfaction scale (Pat-UKU-ConSat)*	Review of literature, Professionals' viewpoint	An extensive literature review, experiences of rating scales, Clinical knowledge of needs of patients as well as of mental health services	0.80	Loevinger's coefficients were 0.36 for the total score and ranged from 0.40 to 0.43 for sub-scores
Jenkinson et al, 2002 (71)	Picker Patient Experience questionnaire (PPE-15)*	Review of literature, Patients' and professionals' viewpoints	Consultation with experts, A systematic literature review, Patient focus groups, In-depth interviews with patients, Item selection based on item analysis.	0.80-0.87	Item-total correlations ranged from: 0.42-0.54 for UK, 0.34-0.58 for Germany, 0.23-0.48 for Sweden, 0.36-0.49 for Switzerland, 0.25-0.57 for USA.
Joyce et al, 2010 (72)	Alberta Continuity of Services Scale-Mental Health (ACSS-MH)*	Review of literature, Patients' and families' viewpoints	A systematic review, In-depth qualitative interviews with severe and persistent mental illness patients (n=36) and their families, Item selection was based on a pre-test (n=52), a pilot-test (n=319) and item analysis	0.72 (0.52-0.72)	Second-order structure (three first-order factors and one second-order factor) All three-first-order factors demonstrated substantial associations

					with the second-order factor (Pearson's coefficients ranged from 0.63 to 0.81, p<0.0001) Inter-subscale correlations ranged from 0.26 to 0.44
Kertesz et al, 2014 (73)	Primary Care Quality-Homeless (PCQ-H)*	Review of literature, Professionals' and patients viewpoints	Identification of constructs derived from IOM's reports, A card-sort ranking exercise with patients (n=26) and clinical providers/experts (n=10), Semi-structured qualitative interviews with patients (n=36) and clinicians/experts (n=24), supplemented by 4 focus groups, Item selection based on a consensus vote and psychometric analysis	0.96 (0.75-0.92)	1 factor Subscales correlations ranged from 0.51 to 0.78
Kolb et al, 2000 (74)	Inpatient Psychiatric Questionnaire*	Review of literature, Professionals' viewpoint	Review of published questionnaires, patient satisfaction literature, Recommendations from inpatient psychiatric administrators and staff, Item selection based on subjective evaluation, item analysis and factor analysis	0.74-0.88	6 factors Inter-scale correlations ranged from 0.35 to 0.58
Larsen et al, 1979 (75)	Client Satisfaction Questionnaire (CSQ-8)*	Review of literature	Review of published and unpublished sources, Item selection based on the items' rating by professionals (n=32) and advisory personnel (n=31), and the results of factor analyses	0.93	1 factor (43%) Inter-item correlations ranged from 0.41 to 0.85
Lelliott et al, 2001 (76) **Blenkiron et al, 2003 (77)**	Carers' and User's Expectations of Services, User Version (CUES-U)*	Review of literature, Patients' viewpoint	A comprehensive literature search of surveys and other instruments, Two focus groups and in-depth semi-structured interviews with service users of mental health (n=7)	NA	3 factors (53% for part A and 50% for part B)
Llyod-Evans et al, 2010 (78)	Camden Content of Care Questionnaire – Patient version (CCCQ-P)*	Review of literature, Patients' and	A review of content of care measures, Qualitative research with staff at an	NA	NA

		professionals' viewpoint	inpatient ward and a crisis house, A Delphi process with community mental health professionals, Consultation with the steering group of The Alternative Study, comprising service users, carer representatives, clinicians and researchers		
MacInnes et al, 2010 (79)	Forensic Satisfaction Scale (FSS)	Patients' viewpoint	Five focus groups with service-users (n=27), Item selection based on factor analysis	0.91 (0.50-0.95)	7 factors
Madan et al 2014 (80)	Menninger Quality of Care (MQOC)*	Professionals' and patients' viewpoints	Informal discussions with unit staff and patients	0.92	4 factors (66.0%) Item to total correlations ranged from 0.36 to 0.67
Mavaddat et al, 2009 (81)	Patient Experience Questionnaire (PEQ)*	Patients' viewpoint	Nine focus groups (n=56), of which six with patients with severe mental illness and three with patients with common mental health problems, Item selection based on lowest item-total correlations	0.94	2 factors (55.6%) Inter-item correlations <0.80 (data not shown)
Mayston et al, 2017 (82)	Mental Health Service Satisfaction scale (MHSSS)*	Review of literature, Professionals' and patients' viewpoints	Review of existing measures, In-depth interviews with service users (n=6), Five focus group discussions including two with services users and three with caregivers (n=40), Four items were added after review by experts from a research consortium	0.92	1 factor (38.9%) Item-test correlations ranged from 0.32 to 0.78 Average inter-item correlation ranged from 0.30 to 0.33
McGuire et al, 2007 (83)	Scale To Assess the Therapeutic Relationship - Patient version (STAR-P)*	Review of literature, Professionals' and patients' viewpoints	Semi-structured interviews conducted with clinicians (n=12) and patients (n=10), Nine existing TR scales was assessed by clinicians and patients for their applicability to community care setting, Item selection based on factor analyses and test-retest reliability	(0.76-0.91)	3 factors

Meehan et al, 2002 (84)	Inpatient Evaluation of Service Questionnaire (IESQ)*	Review of literature, Patients' viewpoint	Eight focus group discussions with inpatients (n=66), Additional service aspects from literature research, Item selection based on patients' importance ratings of items (n=72)	0.95 (0.78-0.93)	3 factors (59%)
Misdrahi et al, 2009 (85)	4-Point Ordinal Alliance Scale (4PAS)*	Review of literature	An extensive literature review, in particular the Helping Alliance Questionnaire from which all items involving psychotherapy aspects were removed and the remaining questions were rewritten	0.91	2 factors (65%) High item-test and item-rest correlations (data not shown)
Moutoussis et al, 2000 (86)	Psychiatric Outpatient Satisfaction Questionnaire*	Review of literature, Patients' viewpoint	Items were derived from the PCSQ instrument, which was adapted to include specific aspects of outpatient care based on a literature review, Input from users through the Community Health Council (CHC)	NA	NA
Nabati et al, 1998 (87)	Satisfaction Index - Mental Health (SI-MH)*	Review of literature	The questionnaire was adapted from the Satisfaction Index by inserting the word 'mental' before 'health care' in each item	0.90	1 factor
Nordon et al, 2013 (88)	Patient Satisfaction with Psychotropic (PASAP)*	Review of literature	An extensive search of the literature, Item selection based on semi-structured interviews with psychiatric outpatients (n=30)	0.85	1 factor (43.4%) Item-total correlations ranged from 0.36 to 0.80 Inter-item correlations ranged from 0.08 to 0.82
Oades et al, 2011 (42)	Consumer Evaluation of Mental Health Services (CEO-MHS)*	Patients' viewpoint	Focus groups and in-depth interviews with consumers, Brainstorming methods, Item selection based on the reliability and usefulness of the questionnaire, followed by a factor analysis	0.92 (0.80-0.92)	2 factors (36.7%) Corrected item-total correlation ranged from 0.53 to 0.73 for factor 1 and from 0.32 to 0.62 for factor 2
Rose et al, 2011 (89)	A short form of Consumer Evaluation of Mental Health Services (CEO-MHS)*	Patients' viewpoint	Inclusion of two new items by a consumer advisory committee, Item selection based on an	0.84	1 factor

			automatic backward stepwise regression analysis		
Ortiz et al, 2012 (90)	Inpatient Consumer Survey (ICS)*	Review of literature, Patients' and professionals' viewpoints	The outpatient Mental Health Statistics Improvement Program (MHSIP) Consumer Survey was used as foundation, A series of meetings with a workgroup including representatives of consumers, MHSIP Policy Group, a research consultant and NRI-BHPMS staff, Item selection based on factor analyses	0.94 (0.73-0.87)	6 factors (62.2%)
Parker et al, 1996 (91)	Patient satisfaction Survey*	NA	Item selection based on factor analyses	NA	4 factors (34.0%)
Pellegrin et al, 2001 (92)	Charleston Psychiatric Outpatient Satisfaction Scale (CPOSS)*	Review of literature, Professionals' viewpoint	Items selected from an inpatient satisfaction scale, A comprehensive review of studies that used surveys and focus groups, Input from psychiatric clinicians	0.87	NA
Perreault et al, 2001 (93)	Opinion Questionnaire on Outpatient Services (OQOS-21)*	Review of literature, Professionals' viewpoint	Literature review, Individual interviews with the administrators of the clinics and group discussions with the staff of each team, Item selection based on factor analyses, Additional items based on patients' comments from interviews about the original version	0.88 (0.50-0.82)	5 factors (56.3%) Item-total correlations ranged from 0.35 to 0.66
Perreault et al, 2006 (94)	Patients' Perspective on Information Questionnaire (PPIQ)*	Review of literature, Professionals' and patients' viewpoints	Review of the literature, Consultations with mental health workers from the outpatient clinics, In-depth individual interviews with psychiatric outpatients (n=8)	0.90-0.91	Item-total correlations ranged from 0.44 to 0.67 for PPIQ-Importance subscale and from 0.49 to 0.75 for PPIQ-Satisfaction subscale
Peytremann-Bridevaux et al, 2006 (95)	Saphora-Psy	NA	Qualitative methods	NA	NA

Phattharayuttawat et al, 2005 (96)	Thai Psychiatric Satisfaction Scale (TPSS)	Review of literature	Review of the literature based on the definition of the Ware's taxonomy of satisfaction and previous literatures, Item selection based on factor analyses	0.96 (0.72-0.93)	7 factors (52.8%) Item to total correlations ranged from 0.88 to 0.93
Priebe et al, 1993 (97)	Helping Alliance Scale (HAS)*	NA	NA	NA	NA
Rofail et al, 2005 (98)	Satisfaction with antipsychotic medication scale (SWAM)*	Review of literature, Professionals' and patients' viewpoints	Systematic literature searches, Clinical expertise, Discussions with patients, Item selection based on pilot study and psychometric analyses	0.91 (0.89-0.92)	2 factors (40%)
Rose et al, 2009 (99)	CONTINUity of care - User Measure (CONTINU-UM)	Patients' viewpoint	Five focus groups with service users (n=32) each meeting twice, Item selection based on a poor conceptual fit and item analysis	NA	NA
Røssberg et al, 2003 (100) **Røssberg et al, 2003 (101)**	Ward Atmosphere Scale – Real ward (WAS-R)	Review of literature	Some items were adopted from an existing scale, Item selection based on item content, item analysis and internal consistency	0.63 (0.54-0.72)	Corrected item total subscale correlation was 0.34 and ranged from 0.28 to 0.41
Ruggeri et al, 2000 (102) **Ruggeri et al, 1993 (124)**	Verona Service Satisfaction Scale - European version (VSSS-EU)*	Review of literature	Items were adapted from three existing service satisfaction scales, Other items were developed to be relevant to the Italian psychiatric care context, Item selection based on results from previous validation studies	0.96 (0.93-0.96) between EPSILON sites	NA
Rush et al, 2013 (103)	Ontario Perception of Care Tool for Mental Health and Addictions (OPOC-MHA)*	Review of literature	A comprehensive literature review of peer-reviewed papers, An environment scan, Item selection based on the review by the project team and the subgroup of key stakeholders	(0.87-0.96)	4 factors (65.5%) Corrected item-total correlations ranged from 0.60 to 0.84 Inter-scales correlations ranged from 0.32 to 0.51
Schalast et al, 2008 (104)	Essen Climate Evaluation Schema (EssenCES)	Review of literature	Items were derived from three existing studies, Item selection based on scale and factor analyses	0.79-0.87	3 factors Corrected item total correlation coefficients ranged from 0.49 to 0.75

Schröder et al, 2007 (105)	Quality in Psychiatric Care (QPC)	Patients' viewpoint	Items derived from the results of a phenomenographic interview study with a sample of patients (n=20), Item selection based on the authors' review (n=3) and on the pilot test with patients (n=6)	0.98 (0.87-0.95) for QPC-1 0.98 (0.85-0.95) for QPC-2	NA
Schröder et al, 2010 (106)	Quality in Psychiatric Care –Inpatient (QPC–IP)		Item selection based on factor analyses	0.96 (0.75-0.95)	6 factors Scales correlations ranged from 0.41 to 0.71
Schröder et al, 2011 (107)	Quality in Psychiatric Care – Outpatient (QPC–OP)		QPC-OP items were adopted from the QPC-IP instrument, Additional items specific to outpatient care were drawn through interviews with patients from a previous study, Item selection based on factor analyses	0.95 (0.65-0.94)	8 factors Inter-scales correlations ranged from 0.30 to 0.69
Shiva et al, 2009 (108)	Inpatient Satisfaction Questionnaire (ISQ)*	Review of literature	A review of the relevant literature and several specific research articles, Item selection based on consensus vote and psychometric analyses	0.81 (0.70-0.79) for F-ISQ scale and 0.85 (0.70-0.80) for C-ISQ scale	4 factors for F-ISQ scale and 2 factors for C-ISQ scale
Slade et al, 2014 (109)	Clinical Decision-making Involvement and Satisfaction - version Patient (CDIS-P)*	Review of literature, Professionals' and patients' viewpoints	A non-systematic scoping review of existing standardized measures, Semi-structured individual interviews with staff (n=5) and services users (n=4), Two focus groups with service users (n=3 and 5), Translation and cultural adaptation into Danish, German, Hungarian and Italian	0.90 (0.87-0.90)	NA
Slater et al, 1982 (110)	Satisfaction with Mental Health Care (SMHC)*	Professionals' viewpoint	A series of meetings between clinical and research staff, Item selection based on a pre-test	NA	4 factors.

			and item analyses		
Speckens et al, 2000 (111)	Reassurance Questionnaire (RQ)*	NA	NA	(0.66-0.83)	1 factor (28-40%)
Svedberg et al, 2008 (113) **Svedberg et al, 2007 (112)**	Health Promotion Intervention Questionnaire (HPIQ)*	Review of literature, Patients' and professionals' viewpoints	HPIQ instrument was constructed using the conceptions, categories and results from two previous qualitative studies based on interviews with patients and nurses, Item selection based on pilot testing with outpatients (n=15), item analysis and factor analyses	0.90 (0.73-0.88)	4 factors (62%)
Uijen et al, 2011 (114) **Uijen et al, 2012 (115)**	Nijmegen Continuity Questionnaire (NCQ)*	Review of literature	A systematic literature review, Patient interviews conducted as part of a previous study, Item selection based on item analysis and factor analyses	(0.82-0.89) (0.86-0.96)	3 factors (70.3-88.8%) Inter-scale correlations ranged from 0.42 to 0.61 Mean inter-item correlation of the subscales ranged from 0.58 to 0.71 Inter-subscale correlations ranged from -0.03 to 0.80
Ul-Haq, 2012 (116)	Service Satisfaction Questionnaire*	Review of literature	Items were derived from existing questionnaires, supplemented by other items addressing specific issues for day hospitals, A copyright authorization was granted by Shropshire's Mental Health NHS Trust for its use in Ireland	NA	NA
Ware et al, 2003 (117)	CONNECT*	Professionals' and patients' viewpoints	Conceptual framework was derived from an earlier ethnographic study, Clinical experience, Item selection based on cognitive interviews (n=9) and two pilot studies (n=41)	Five CONNECT scales had a Cronbach's alpha >0.80 and the remaining scales were 0.70 or above	Scale to scale correlations ranged from -0.11 to 0.52

Webster et al, 2012 (118)	Patients Satisfaction Questionnaire*	Review of literature, Professionals' and patients' viewpoints	A literature review, In-depth interviews of hospital staff, Focus groups of patients and nursing staff, Additional items generated from the results of the nurses' interviews and patients' focus group	NA	NA
Wongpakaran et al, 2013 (119)	Group Cohesiveness Scale (GCS)*	Review of literature	Items were adopted from the short form of the Engaged subscale of the Group Climate Questionnaire and of the Cohesion subscale of the Therapeutic Factors Inventory	0.87	1 factor Item-total correlations ranged from 0.497 to 0.752
Woodring et al, 2004 (120)	Penn State Inpatient Psychiatry Satisfaction Survey (PSIPSS)*	NA	Item selection	0.94 (0.86-0.92)	2 factors (61.58%)
Wright et al, 2006 (125)	Survey of Health Care Experiences of Patients (SHEP)*	Review of literature	Dimensions were derived from instruments developed by the Picker Institute, Previous qualitative work based on focus groups of veteran patients and their families	(0.59-0.81) for inpatient survey (0.35-0.89) for outpatient survey	NA
Zendjidjian et al, 2015 (121)	Satisfaction with Psychiatry Care Questionnaire-22 (SATISPSY-22)*	Patients' viewpoint	Face-to-face semi-structured interviews with psychiatric inpatients (n=80) on the last day of their stay, Item selection based on statistical analyses and the expertise of the steering committee	0.70-0.95	6 factors (78%) Item-internal consistency ranged from 0.72 to 0.94 Item discriminant validity ranged from-0.02 to 0.64
Zimmerman et al, 2017 (122)	Clinically Useful Patient Satisfaction Scale (CUPSS)*	Review of literature, Professionals' viewpoint	Clinical experience, A review of studies of patient focus groups, surveys that identified the best indicators of quality care, and studies of satisfaction with facets of treatment and clinician behavior, Item selection based on item analysis	0.92- 0.93	Item-total correlations ranged from: outpatient sample: 0.65(0.42-0.76) partial hospital sample: 0.66(0.52-0.76)

* Items are available in the article.

[a] Internal consistency measured by Cronbach's alpha at the scale or dimension level.

Abbreviation: NA, not available.

Table 3. Evaluation of the item generation process of the instruments.

Reference	Pre-study hypothesis	Actual content area	Item identification	Item selection	Unidimensionality	Choice of response scale
Aloba et al, 2014 (43) **Anderson et, 1990 (44)**	✓ The objective was clearly detailed; however, patients were recruited from only one tertiary health care facility and the sample was tilted to patients with schizophrenia (63%).	✓ Face validity was partially supported by the item generation process that included patients. However, the measure was not pre-tested with the target population before being applied on a larger scale.	✓✓ The initial pool of items was generated after reviewing existing instruments and from the results of interviews with patient groups and various health care providers.	✓✓ In the original version, 25 items were developed. A few them were excluded after item analysis because they had too low variance or correlations lower than 0.40 with the sum of the other items. Two other items were also deleted due to a lack of conceptual clarity. Ceiling effects were not a problem with less than 3% of respondents concerned. The resulting 11-item version was tested on a sample of psychiatric patients and showed weak psychometric properties.	✓✓ All items had a factor loading greater than 0.40 and no cross-loading. Cronbach's alpha was modest for the total scale (0.68) as well as for one of the two domains.	✓✓

				However, the deletion of any of the scales' item did not significantly improve its overall internal consistency, and as a result, all items were retained.		
Atkinson et al, 2004 (45)	✓ The TSQM instrument is a generic instrument of patients' experience with their medication that was tested in a sample of patients with various types and forms of medications, as well as various chronic conditions.	✓✓ In-depth interviews with a sample of patients (n=17) were conducted to select the most relevant items from an initial set of items itself developed through an extensive literature review and the input from three patient focus groups (n=30).	✓✓	✓✓ After only keeping the most relevant items for all illness groups using in-depth interviews with patients (n=17), the remaining items were subjected to multistep exploratory factorial analyses. Only the items with the highest factor loadings were retained for inclusion in the final TSQM scales. Additionally, the score distributions were characterized as expected by ceiling effects and skew due to item	✓✓ All items retained in the final version had a factor loading greater than 0.40. Cronbach's alpha was >0.70 in all domains.	✓✓ Two scaling methods were considered for use in the final instrument (VAS and Likert scale). The Likert scaling method proved to be more efficient than the VAS and provided a better distributional characteristics and lower measurement error. As a result, this method was associated with a higher proportion of meaningful variance across a variety of parametric analyses.

				relevance and a continuous self- and clinical selection process.		
Baker, 1990 (46)	✓ The CSQ instrument is obviously reliable under the conditions of this study and there are grounds for being optimistic that future studies will confirm validity. Patients were recruited from different general practices.	✓✓ The instrument was refined so that patients could understand the items and obtain a wide range of opinions. Further, the difficulty of answering the items was assessed by examining additional comments written by patients in the questionnaires. Each of the revised versions was tested by a group of patients.	✓✓ The questionnaire was developed from a review of other existing patient satisfaction questionnaires, supplemented by discussions with fellow general practitioners and patients' personal experience. Two open-ended questions were included in the first version to ensure that all areas were covered by existing statements.	✓✓ Several methods were used to evaluate the statements to be kept: the meaning of the items examined by professionals, the skewness of the responses and the wording of the items for ambiguity. The exclusion of items was also based on the findings of a factor analysis.	✓✓ All factor loadings were greater than 0.40. Cronbach's alpha was 0.91 for the complete questionnaire and greater than 0.70 in all domains, except for items related to overall satisfaction which were used as a separate scale (0.67).	✓✓ A 5-point scaling method was chosen because it was already used in other studies and is easily understood by respondents.
Barker et al, 1999 (47) **Barker et al, 1996 (48)**	✓ The PCSQ instrument was designed and validated in inpatient settings, but it can easily be adapted to outpatient departments or the day hospital, which	✓ The view about the usefulness of the questionnaire was obtained from potential users, with 68% of mental health professionals and 64% of inpatients (or MIND workers) responding	✓✓ The instrument was developed using a combined approach that included a literature review and the perspectives of patients and professionals.	✓ The reliability and some aspects of the validity of the PCSQ instrument were examined and all the items were kept in the final version. However, data were not analyzed	✓ Cronbach's alpha >0.70. No factor analysis was performed.	✓✓

	would require further psychometric testing with this target population.	positively. In addition, 64% and 71% respectively considered that all the areas covered by the questionnaire were important and 53% and 64% felt that other questions should be added.		in terms of missing data or ceiling and floor effect.		
Berghofer et al, 2011 (49)	✓✓ The ECS instrument is intended for use with people with chronic mental illnesses in community treatment settings. Half of the study sample consisted of people in assisted outpatient treatment, but there were no statistically significant differences with the other participants, except for psychiatric diagnosis.	✓ It is specified that a particular emphasis was placed on understanding the items by patients with chronic mental illnesses. However, no evidence was provided regarding the evaluation of the measure with the target population prior to larger scale administration.	✓The ECS items were drawn from an existing questionnaire that was translated and submitted to an expert panel to improve its wording and to adapt it to the US healthcare context. The original measure itself was developed from semi-structured face-to-face interviews with service users from outpatient settings.	✓✓ The item selection process took into account the proportions of missing values, a factor analysis and the assessment of reliability at baseline. Two items were excluded because they failed to have sufficient loading (>0.50) on any of the factors and another item because it did not statistically increase scale reliability. Ceiling and floor effects were not examined.	✓✓ All items included in the final version had loadings of >0.50 on at least one factor and Cronbach's alpha was >0.70 for the total scale and in all domains.	✓✓ Use of the same scaling method as the original WPI instrument.

Bjertnaes et al, 2015 (123)	✓✓ The POPEQ-OS instrument can be used to assess the experiences of psychiatric inpatients on-site, but further research is needed to evaluate its usefulness as basis for external quality indicators.	✓✓ The questionnaire was pre-tested through cognitive interviews with psychiatric inpatients (n=10). Overall, the measure was found to be useful and relevant to the patient group.	NA The PIPEQ-OS items were adapted from an instrument originally developed to measure the experience of psychiatric inpatients after their discharge from hospital, leading to many missing data. This new version was developed to measure the patient experience on site and some items have been reformulated accordingly. Moreover, adjustments have been made to ensure that this revised version was consistent with the latest developments of the national program.	✓✓ Some items were discarded due to a high rate of missing or "not applicable" responses (>20%) and a significant ceiling effect (>50%).	✓✓ All items had a factor loading >0.40 and all scales met the criterion of 0.7 for Cronbach's alpha.	✓✓
Blais et al, 2002 (50)	✓✓ The PEC-5 instrument is a brief multi-faceted services evaluation tool and program that has adequate psychometric properties for effective program evaluation.	✓✓ Staff and patients rationally reviewed the initial pool of items, leading to the deletion of some items that were not relevant to the concept to be measured.	✓✓ The items were developed from a literature review and qualitative data gathered from focus groups with patients and unit staff.	✓✓ The original item pool was first reduced based on a rational review and empirical analyses. Subsequently, remaining items with the highest factor loading on the first factor were subjected to a	✓✓ Items were subjected to a Rasch analysis and then a factor analysis that confirmed the unidimensional nature of the scale. Cronbach's alpha was >0.70.	✓✓

				Rasch analysis to produce the PEC-5 instrument.		
Bramesfeld et al, 2007 (52)	✓✓The applicability of the concept of responsiveness to mental health care was assessed through qualitative work and the instrument thus developed provides a picture of inpatient and outpatient mental health care performance. Future work should make it possible to revise and shorten the instrument to make it short, self-administered and easy to understand.	✓ Patients were involved in qualitative work to assess the applicability of the concept of responsiveness to mental health care.	✓ The Multi Country Service Study (MCSS) questionnaire was developed by WHO to assess the responsiveness of health systems. It was developed from a broad-scale literature review concerning patient satisfaction and quality of care, as well as a meeting of experts. Its German version was then tailored to suit mental health care context based on previous qualitative work evaluating the service users' expectations. As a result, the terminology was adapted and a “continuity” domain was added, as well as a section evaluating experiences with day care and hostel care.	X	X The structure and psychometric properties of this version were not examined.	✓✓ The original MCSS questionnaire is organized into separate modules, one of which is specifically dedicated to mental health and substance use. In accordance with the original instrument, the distance or time duration items are expressed in numerical values, the report items used a 4-point scale and the rating items used a 5-point scale.

Brunero et al, 2009 (51)	✓✓ From the survey results, interventions have been developed, which are hoped to further improve the service and enhance consumer outcomes.	✓✓ The draft questionnaire was first reviewed by a group of senior mental health professionals for judgement on relevance and applicability and then subjected to a series of discussion groups (n=3) with consumers.	✓✓ The questionnaire was developed in partnership with consumers and health staff through a series of discussion groups. It was built to reflect national mental health standards and contemporary patient satisfaction literature.	X	X	✓
Bruyneel et al, 2018 (39)	✓✓ The instrument was tested in different types of mental health care organizations and patient groups.	✓Face validity was ensured by the generation process of the measure which included patients at all stages. In addition, special attention was given to patient representatives' suggestions on clear wording of the items. But no cognitive debriefing phase was undertaken.	✓✓ The items were generated through a scoping literature review, followed by focus group-type brainstorm sessions with patient representatives (n=6) and mental health professionals (n=18). The items were then iteratively selected using two Delphi-rounds with professionals (n=60 and n=52), followed by patient representatives (n=48) selected the most relevant and important items.	✓✓ The most relevant items were selected from the initial pool by mental health professionals and patient representatives. All items submitted to the psychometric evaluation phase were retained at the end of the tests. The number of missing data ranged from 8.4% to 33.5% between dimensions. Ceiling and floor effects were not	✓ The dimensionality of the instrument was examined using exploratory structural equation modelling and bifactor model. Both models were shown to be ideally fit for a nine-factor solution. This was then corroborated from a confirmatory factor analysis using a holdout sample. Some items showed low loadings (less than 0.30) or high cross-loadings. However, these items have been retained due to their strong association with overall ratings and should be further discussed in the development of future	✓All items had 4-point response categories, except for two items related to the discharge management domain which were scored with a yes/no response format. The overall rating items used an 11-point scale. No justification was provided for the use of such a variety of rating scales.

				examined.	versions. Cronbach’s alpha was not computed.	
Caruso et al, 2013 (53)	✓ The objective and target population have been clearly detailed. The limited sample size, from only one region of Italy, is one of the study limitations to be taken into account for generalization of the results. Also, the measure was validated using a limited range of group treatments.	✓✓ The draft questionnaire was carefully reviewed at a meeting with 12 patients and all items were deemed understandable and valuable.	✓ Patients were not included in focus group discussions to generate the item pool, but were subsequently involved in modifying the content to make it clear and easy to understand.	✓✓ The distribution of the data was examined in terms of skewness and kurtosis. Low correlating items were not eliminated because it did not substantially improve the explained variance of the model, unlike the highly skewed items. Next, several principal component analyses were carried out. Four highly loaded items (>0.50) on each component were kept.	✓✓ All items were highly loaded (>0.50) on each component. Cronbach’s alpha was >0.70 in all domains.	✓✓
Clement et al, 2012 (16)	✓ Caution should be taken due to patient recruitment: the sample was self-selected and restricted to	✓✓ The draft version was submitted to an expert panel, including people with mental illness,	✓ The initial pool of items was derived from a literature review to avoid duplication of effort. Additional items were then added from	✓First, some barriers items were deleted and merged by the research team according to their content	✓ Only Cronbach's alpha for the treatment stigma subscale was computed (0.89). No factor analysis was performed.	✓✓

	participants with internet access. Also, one of the recruitment pathways was the website of an anti-discrimination programme, which can result in a biased sample.	who provided feedback. Two items were added following this expert panel's suggestion. A free-text comments section was added to capture the overall opinion on the questionnaire. Additionally, the measure was broadly positively evaluated by participants, with a median overall evaluation rating of 8 on a 10-point scale.	the research team's knowledge of aspects considered important in the qualitative and theoretical research literature. Feedback from an expert panel was then used to refine the draft, and the psychometric findings, as well as further discussion within the research team, produced the final version.	analysis. Other items were eliminated following feedback of an expert panel. After the psychometric analyses, other items were discarded because they were highly inter-correlated (>0.70), or because less than 10% of respondents endorsed them to be a major barrier or due to a large conceptual overlap. Additionally, some items were amalgamated. There were no missing data due to the design of the online survey.		
Eisen et al, 2001 (54)	✓✓ The MHSIP survey reliably assessed treatment outcome.	✓✓ Consumers from three states were interviewed and confirmed that the instrument addressed important concerns, was easy to	✓✓ Input from a wide range of stakeholders was involved in the generation process of the measure.	✓ All initial items were kept in the final version.	✓ All items had a factor loading greater than 0.40 except for two items that did not fit well with the factors identified. However, these items were kept. Cronbach’s alpha was >0.70	✓✓

		understand, and was easy to complete.			in all domains.	
Eisen et al, 2001 (54) **Eisen et al, 1999 (55)**	✓✓ The CABHS survey reliably assessed features of the insurance plan	✓✓ After carefully review by experts, a revised version was tested in consumer focus groups to ensure that all the important domains were included and that all items were understandable.	✓ Input from a wide range of stakeholders was involved in the generation process of the measure.	✓ All initial items were kept in the final version.	✓✓ All items had factor loading >0.40 and internal consistency for four of the five domains was acceptable (Cronbach's alpha ≥ .70).	✓✓ CABHS responses options are tailored to each set of questions. These rating scales are the same as those used by the CAHPS survey which was used as the framework for the development of the CABHS.
Eisen et al, 2002 (56)	✓ The study sites were healthcare organizations that chose to use the PoC survey and it was not possible to determine the percentage of discharged patients who completed the survey. These findings may have limited the representativeness of the sample.	X	X The items were generated exclusively from a literature review of existing measures.	✓✓ Data analysis included frequency distributions of reports and ratings, a factor analysis, regression analyses and the general linear model least-squares regression procedure. Surveys with more than two missing items were omitted from the analyses. For surveys with two or fewer missing items, missing data were imputed.	✓ Exploratory and confirmatory factor analyses were performed to determine the underlying structure of the instrument (but factor loadings not shown). Cronbach's alpha did not reach the recommended cut-off in two of the four domains.	✓ No justification was provided for the use of a variety of rating scales.

Eton et al, 2017 (58) **Eton et al, 2015 (57)**	✓ The PETS instrument allows for comparisons across a wide variety of healthcare settings and patient conditions. This cross-sectional study provided initial evidence of reliability and validity in a diverse patient sample and should be replicated in a large-scale prospective validation study.	✓✓ First, the draft measure was reviewed by a stakeholder panel and some revisions were made accordingly these feedbacks. Second, the remaining draft was submitted to cognitive pre-testing with patients in two rounds (n=11 and n=12). A number of items were removed due to low importance, frequency of endorsement, relevance, or redundancy. In addition, several modifications were made to improve the scale and to clarify meaning.	✓✓ The questionnaire was developed from a conceptual framework itself based on qualitative interviews with a sample of patients (n=50), followed by four patient focus groups (n=25) to test the relevance of the conceptual framework and identify unrepresented issues.	✓✓ First, a number of items were deleted after review by a stakeholder panel and then after the cognitive pre-testing phase. Then, other items were deleted during the descriptive analyses due to a large number of missing data (≥25%), strong correlations (≥ 0.80), and a lack of conceptual adequacy or when an item had strong inter-domain correlations. Also, descriptive statistics, including floor and ceiling effects and skewness coefficient, were computed. Overall, scale scores were positively skewed toward lower burden, with substantial numbers	✓✓The factor structure of the instrument was evaluated using a CFA based on the initial conceptual framework and subsequent modifications to obtain a more parsimonious model. The resulting 9-factor model provided a satisfactory fit to the data. Cronbach's alpha was >0.70 in all domains.	✓ Two response scales were used depending on the item content. No justification was given for the choice of using several scoring methods.

				scoring at the floor.		
Evans et al, 2012 (59)	✓✓ The VOICE instrument is a psychometrically robust measure which encompasses the issues that service users prioritize and is both acceptable and accessible to people with a range of diagnoses and severity of illness for acute settings.	✓✓ Service users were fully involved in all stages of the development process of the measure through a participatory methodology. Expert panels considered that the measure was of an appropriate length and breadth. In addition, participants in the feasibility study felt that the measure was comprehensive and understandable. VOICE therefore reflected the issues that service users considered most important.	✓ VOICE instrument was generated exclusively from the service users' viewpoint through repeated focus groups. The draft measure was then reviewed by expert panels, themselves consisting of service users. Optional free-text sections were included to capture additional qualitative data to support content validity.	✓ Some items were removed for their relevance and to prevent duplication, while others were deleted due to their poor reliability. However, ceiling and floor effects, as well as the rate of missing data, were not examined.	✓ Cronbach's alpha was high for the total VOICE scale (0.92). No factor analysis was performed.	✓✓
Forouzan et al, 2014 (60)	✓ This study reported the feasibility, reliability, and validity of the MHSRQ instrument used to assess mental health	✓✓ The new questionnaire was developed following a rigorous translation process in which the back-translated version was modified by an	✓ The questionnaire was developed from an existing instrument that was tailored according to the findings of previous qualitative studies on the expectations of services	✓✓ Three items related to the domain of access did not perform well due to their different formulation, but they were	✓ As the factor structure of the original instrument was already established, the new questionnaire was only subjected to a confirmatory factor analysis. All items loading was >0.30, expect for three items related to the	✓ Distance or duration items were expressed in numerical values, items related to the quality of basic amenities and rating questions used a 5-

	system responsiveness in Iran. Unlike the original instrument, inpatients were not included due to difficulties in accessing this group.	expert panel until it was comparable to the original English version. Two bilingual experts then independently confirmed this translation. Additionally, the approved Farsi version was then tested in a pilot study with outpatients (n=20) and some items were revised for clarification.	users and providers. As a result, some changes were made, such as adding a new domain of effective care, as well as some questions to existing domains. Additionally, some domains were merged and other divided.	nonetheless kept in the final version. Modification attempts were made and, consequently, two items were deleted because they had high residuals (>0.80). The results showed a satisfactory feasibility since the item missing value was lower than 5.2%. Floor/ceiling effects were not examined.	access domain. Cronbach's alpha reached the acceptable cut-off of 0.70 for six of the eight domains.	point scale with different response categories. The remaining items used a 4-point scale.
Garratt et al, 2006 (61) **Olsen et al, 2010 (62)**	✓✓ The POPEQ instrument is brief and was designed to be relevant to all psychiatric outpatients.	✓✓	✓✓ POPEQ instrument was developed based on an extensive literature review of questionnaires used in psychiatry and qualitative interviews with patients (n=12) to ensure that all items adequately address important aspects of patient experience. Thereafter, the draft version was discussed within a network of clinical staff and pre-	✓✓ The level of missing data was low for both surveys (<5%). As expected, mean item scores were skewed towards positive experiences. However, all items were kept following analyses using either classical test theory or Rasch analysis.	✓✓ Factor analysis showed factor loadings greater than 0.40 on a single factor (ranging from 0.54 to 0.84). Cronbach’s alpha was >0.70.	✓✓

			tested with patients (n=1238) to produce the final version.			
Gensichen et al, 2011 (63)	✓✓ This German version of the PACIC instrument showed good psychometric properties in a sample of patients with major depression.	✓✓ The questionnaire was pilot tested with a separate sample of patients (n=130). Only items that patients had no difficulty understanding were retained.	✓ The initial set of items was generated by a national pool of experts on chronic illness care and the Chronic Care Model framework.	✓✓ Firstly, the items that were selected from the larger pool were those that showed adequate variability, were not difficult to understand and that best represented the underlying concepts. There were possible ceiling effects for two subscales (8.9-12.9%), as well as floor effects for one subscale (4.6%). The proportion of missing values ranged from 0.7 to 5.4%. All items submitted for analysis were retained at the end of the tests.	✓All but two items had a factor loading greater than 0.40. For items that did not have a clear loading pattern, they were assigned to a factor after analyzing their content. Cronbach's alpha was good for the total scale (0.91), but not for one of the five subscales (0.45).	✓✓
Gigantesco et al, 2003 (64)	✓✓ The ROQ-PW questionnaire is an adequate tool for	✓ To assess face validity, the draft version was	✓✓ The initial pool of items was generated following a literature review and an open-	✓Of the 20 initial statements, only half were kept in	✓✓ All items submitted to the tests had a factor loading >0.40. Cronbach's	✓✓ The use of a 5-point Likert scale with a range of responses

	evaluating patients' opinions on the care provided in inpatient psychiatric wards. Also, it could be slightly modified for use in other settings, such as day centres, residential facilities and day hospitals.	submitted to two focus groups with psychiatric inpatients (n=15) to capture their opinions on the relevance, usefulness and clarity of each item. Patients found that the questionnaire was clear and acceptable, but noted that there should have been items on the quality of food and the severity of the side effects of drugs. In addition, after performing the questionnaire validation, two items have been added to cover missing content areas.	questionnaire survey with a sample of inpatients (n=40). The preliminary version was then submitted to psychiatrists (n=8) to assess its content validity. All psychiatrists felt that ROQ-PW instrument was a useful measure.	the final instrument to maximize acceptability and improve the potential response rate. Item selection process was made according to the items that were ranked most important by psychiatrists, according to patients' ratings. Data distribution was not analyzed in terms of missing data or ceiling and floor effects. All these items were kept at the end of the tests and were supplemented by two additional items, but the psychometric evaluation was not replicated.	alpha was good for the total ROQ-PW scale (0.82) but not satisfactory for two of the three domains.	wider than "satisfied/dissatisfied" allows patients to express a positive opinion while highlighting specific needs for improvement.
Glick et al, 1991 (65)	X	X	X	X	X	✓✓
Hansson et al, 1995 (66)	✓✓ The instrument was tested in pilot	✓✓ Patients were asked to what extent	✓✓ The development process used an existing	X Ceiling and floor effects and missing data were not	✓ Cronbach’s alpha was good for both versions. No	✓✓ The wording of the response

studies using large and multicentre samples. The questionnaires have proven useful in both inpatient and outpatient settings and in samples covering the full range of psychiatric disorders.	these questionnaires covered all aspects of care considered relevant and important to them. To do this, unstructured interviews (n=77 and n=94) followed by a ranking of the identified treatment characteristics by order of importance (n=78 and n=84), were conducted. It appeared that dimensions of care given the highest priority by patients were well covered in the two questionnaires . On the other hand, the patients were given the possibility of making open-ended comments on which items were difficult to understand or answer, and also to identify if some aspects of care were not covered by the	questionnaire as a starting point combined with the views of patients, relatives and professionals.	reported. Based on the pilot studies, the questionnaires were revised and some items were discarded, but the process was not described.	factor analysis was performed.	alternatives differed slightly according to the items' content.

		questionnaire.				
Hester et al, 2015 (67)	✓ The SEQUenCE instrument was developed and validated with service users with a diagnosis of either bipolar disorder or a psychotic disorder and performed well in tests of reliability and validity in a secondary mental health service context. Nevertheless, recruiting from an independent mental health service may have limited the the generalizability of the results to users of the public health system. The analyses should be extended to other psychiatric settings and populations.	✓✓ The acceptability and face validity of the instrument was evaluated with a sample of patients (n=10). Participants noted the importance of each item on a 5-point Likert scale. Items that did not reach the mean of 4 were excluded from the draft. In addition, participants answered standardized questions on the layout, understanding and completeness of the instrument. Overall, patients found the instrument easy to understand and that all items were moderately important components of the quality of mental health care.	✓✓ The items were generated from qualitative analysis of data gathered from patient focus groups (n=29) and analysis of clinical practice guidelines due to their widespread use in many healthcare systems. Items from both sources were compared and the draft was refined accordingly, until there were no more redundant or irrelevant items.	✓ During the instrument's development, the items generated by the two sources of information were carefully compared and some of them were merged or deleted. Other items were then excluded when assessing face validity and acceptability, as well as in the field-testing, resulting in a 40-item instrument. The percentage of missing data was low (0.3%) and no pair of items had a correlation score of >0.80, except one (0.81). Average quality of care scores from field-testing was quite high.	✓ Cronbach's alpha was good (0.87), but no factor analysis was performed.	✓✓

Howard et al, 2001 (68)	✓✓ The KY-CSI instrument is a measure of consumer satisfaction, designed with consumer involvement. Future research efforts will focus on replicating this study at other hospitals sites, to improve instrument validity and reliability, including development and testing of a self-administered version of the KY-CSI.	✓✓ The KY-CSI instrument was modified based on feedback from two clinician panels and field testing with psychiatric inpatients (n=18).	✓✓ Items were developed using a combined approach: a literature review and the perspectives of patients and professionals.	✓✓ Mean substitution was used, resulting in a complete set of responses for all items submitted to psychometric analyses. Ceiling and floor effects were not examined. At the end of the test, the 19 items initially developed were all retained.	✓ All items exhibited factor loadings of at least 0.35 (data not shown). Cronbach's alpha for the KY-CSI total scale was 0.90 but it was slightly less than 0.70 for one of the three subscales (0.69).	✓
Ivarsson et al, 2007 (69) **Ahlfors et al, 2001 (70)**	✓✓ The results showed that the patient self-rating scale gave results comparable to those obtained by independent assessors and was therefore suitable for use in ordinary clinical practice.	X These self-rating version seems to have captured the content of the original instrument. However, no evidence of face validity was provided for either version.	X Pat-UKU-ConSat is the self-reported version of an original instrument designed to be administered by independent professional interviewers. Some changes were made to make the scale easily understandable and consumer-friendly. Additionally, one	✓ Missing data, as well as the characteristics of the response distribution were not examined. But results showed that this self-rating version was psychometrically sound and consequently all the items tested	✓ Unidimensionality was evaluated using a non-parametric Mokken analysis based on the Loevinger coefficient of homogeneity. Coefficients for sub-scores were acceptable (≥0.40), while it was only just acceptable for the total score (0.36). Cronbach's alpha was good (0.80).	✓✓

			screening question was added to deal with patients without prescribed medication. Patients were not included in the development process.	were kept.		
Jenkinson et al, 2002 (71)	✓ The PPE-15 instrument was designed to be easily and quickly completed by patients, and to enable straightforward scoring. It provides a basic set of questions that should be applicable in all hospitals and relevant to all patients.	✓✓ The original US version was assessed through in-depth interviews with patients, then redrafted and piloted several times to produce the final questionnaire in its long form. The translated versions were then tested with patients in the different countries to test comprehension as well as cultural and linguistic relevance.	✓✓ PPE-15 instrument is a shorter version of an existing questionnaire that was initially developed from consultation with experts, a systematic literature review, patient focus groups and in-depth interviews with patients.	✓ Items were deleted because they were not applicable to a large proportion of respondents or because their removal resulted in an increase in the reliability of the instrument.	✓ Cronbach's alpha were >0.80 for each country participating in the study. No factor analysis was performed.	✓ The response scale was different depending on the items, but no justification was given about the variety of scales used.
Joyce et al, 2010 (72)	✓ The ACSS-MH is a promising instrument for evaluating continuity in specific programs or sectors of mental	✓✓ The draft questionnaire was first reviewed by researchers and active clinicians for relevance and readability. This	✓✓ The attributes of perceived continuity of care were identified from qualitative data obtained through interviews with patients and their families as also	✓✓First, some items were removed after a pre-test phase followed by a pilot test to produce the draft submitted for	✓✓ All retained items had a factor loading >0.40. However, Cronbach's alpha was acceptable for the second order factor, but two first-order factors had an	✓✓

	health systems. Nevertheless, ACSS-MH has yet to be assessed among persons from all points on the severity continuum, as the sample consisted primarily of SPMI patients. Also, the measure was developed and evaluated in the context of a publicly funded health care system.	panel of items was then pre-tested with 52 patients, family members and mental health professionals. The subsequent revised version was then pilot tested with 319 patients with SPMI and some changes were made accordingly before psychometric evaluation.	from a review of the literature.	psychometric evaluation. Second, other items were excluded because they did not apply to a large number of respondents and thus the rate of missing data was high. Finally, other items were excluded because the representation of response categories was insufficient.	alpha <0.70.	
Kertesz et al, 2014 (73)	✓✓ Systematic application of qualitative and quantitative methods supported the development of a brief patient-reported experience questionnaire.	✓✓ Patients were involved in the item generation process so that the instrument covered all patients' priorities and concerns in the specific context of homelessness. All items were reviewed by a multidisciplinary team. In addition, the draft measure was subjected to cognitive interviews	✓✓ The items were generated from a review of the literature associated with the perspective of patients and clinicians/experts from this specific setting.	✓✓ Firstly, the most relevant items were selected by a consensus vote within a multidisciplinary team. Items were then deleted due to a high rate of missing data (>10%), by the merging of highly correlated sub-scales (>0.80) identified during CFA and following analyses guided by	✓✓ Unidimensionality was approached by subjecting items to a preliminary confirmatory factor analysis, then performing IRT models on sub-sets of items identified as loading on individual factors (data not shown). Cronbach's alpha was >0.70 for total scale and in all domains.	✓✓ A 4-point Likert scale was used to simplify administration in resource-poor environments with low-literacy populations.

		(n=12) and resulted in only slight wording changes.		the item response theory.		
Kolb et al, 2000 (74)	✓✓ The findings of this study are drawn from a convenience sampling, which may question the generalizability of these data beyond the participating facilities. However, these results were cross-validated using responses collected post-study from facilities that were independent of the original sample.	✓✓ The instrument was reviewed by a panel (n=109) including professionals (administrators and staff), patients and family members to assess the importance, feasibility and applicability of each item. Items that did not meet the desired requirements were excluded.	✓ Items were developed from a literature review and recommendations of professionals.	✓✓ The items were first selected based on the subjective evaluation made by the panel members. Other items were then deleted after item analysis due to high inter-item correlations (>0.80), skewness or kurtosis, missing data rates, etc. Finally, items were excluded because two of them did not sufficiently load on any of the identified factors.	✓✓ All items had a factor loading greater than 0.40 and no cross-loading. Cronbach's alpha was high.	X
Larsen et al, 1979 (75)	✓ The CSQ-8 is a scale to assess general client/patient satisfaction with services. The study sample was exclusively outpatient, but included persons	X The instrument was not pre-tested on a small sample of patients to assess the relevance and understanding of the items before its larger scale administration.	X The items were derived from a review of published and unpublished literature to identify the potential determinants of satisfaction with services.	✓✓ First, the initial pool of items was reduced by half based on the items' ratings by professionals (n=32) and then by advisory personnel (n=31). The resulting reduced	✓ Factor analysis was performed and only items with the highest factor loadings were selected (data not shown). Cronbach's alpha was high (0.93).	✓✓ Each item was scored with a 4-point Likert scale without the neutral position.

	receiving a variety of treatments.			version was then pilot tested on a sample of outpatients, and only items that were highly loaded on the first factor and which exhibited good inter-item and item-total correlations were retained. The distribution of responses was skewed. Missing data were not analyzed.		
Lelliott et al, 2001 (76) **Blenkiron et al, 2003 (77)**	✓✓ The development and testing of CUES-U have demonstrated the feasibility of applying a self-rated measure of the expectations and experience of users of mental health services.	✓✓ The draft questionnaire was first modified in light of comments from an advisory group of service users. Subsequently, in a pilot study involving service users (n=82), structured feedback was provided on language, coverage of relevant domains, clarity, length and	✓✓ The CUES-U instrument covers all aspects that patients have identified as being their priorities using focus groups and interviews, combined with a comprehensive literature review.	✓✓ The average rate of missing data was 5% and ranged from 2.4% to 10% for Part A and from 1.8% to 9.2% for Part B. The highest response categories were endorsed by 76.6% and 73.6% respectively for Parts A and B. These high levels of the satisfaction scores could lead to	✓ All items had a factor loading > 0.40. Although some items exhibited cross-loading, all were kept in the final version. Cronbach's alpha was not calculated.	✓✓

		acceptability of the instrument. All indicated that the CUES-U instrument was generally clear and acceptable, and 71% also found it comprehensive.		a "ceiling effect" masking any potential differences between patients, but this issue was not further examined. All the items submitted for analysis were retained.		
Llyod-Evans et al, 2010 (78)	✓✓ This study proposes a multi-method approach to measure the intensity and nature of care in mental health inpatient services. The development is described for four instruments measuring different perspectives. However, uncertainty about the psychometric properties of these instruments makes interpretation of divergence in results from the measures	✓✓ The instrument was piloted both with staff and patients to ensure its completeness, acceptability and clarity for respondents. In addition to filling the measure, feedback was obtained. Some changes were made according to these findings. The subsequent revised version was then submitted to the steering group of The Alternatives Study and no further amendments were	✓✓ Information sources and measurement methods were drawn from the results of a literature review of measures of the content of mental health care. Items were then derived from three complementary sources: a Delphi process with mental health professionals, a qualitative study with staff and suggestions from the steering group of The Alternative Study. This advisory group included service users, carer representatives, clinicians and	X Psychometric properties could not be evaluated due to the lack of directly comparable instruments. As a result, the overall conclusion was about the uncertainty surrounding the robustness of the instrument.	X	✓✓ The intensity of care was assessed using a 7-point Likert scale, while the nature of care used a yes/no response format.

	problematic.	made.	researchers.			
MacInnes et al, 2010 (79)	✓✓ The FSS instrument can be used in low and medium secure inpatient settings to assess service-user satisfaction with forensic mental health services.	✓✓ The draft version was pre-tested with eight focus group participants before piloting. Ease of use, clarity and level of agreement with the content were assessed and some modifications were made accordingly. Subsequently, participants were also asked to comment on the appropriateness and difficulty of interpreting the items. As a result, a "not relevant" response category was added.	✓✓ Data were generated from service user comments collected through focus groups. The identified themes were then discussed again with the focus groups' participants to enhance the instrument validity.	✓ The initial pool of items was reduced using a factor analysis to produce the 60-item version that was then submitted for psychometric evaluation. No items were discarded at the end of the tests. Ceiling and floor effects, as well as the rate of missing data, were not examined.	✓ Cronbach's alpha was high for the total scale (0.91) and in 4 out of 7 dimensions (>0.70). Two other dimensions reached a moderate reliability (>0.50). A factor analysis was performed to reduce the initial pool of items (data not shown).	✓✓
Madan et al 2014 (80)	✓✓ Preliminary psychometric analyses revealed that this was a reliable and valid measure of patient satisfaction.	✓	✓✓ The instrument was developed using findings from informal discussions with unit staff and patients.	✓ None of the items submitted for psychometric analysis were deleted at the end of the tests. However, ceiling or floor effects, as	✓✓ All items had factor loadings above the recommended 0.40 cut-off and there was no cross loading. Cronbach's alpha was good (0.92).	✓✓ A 4-point metric was chosen in favor of a 5- or 7-point Likert type scale because midpoint (3=neutral) responses and increased response choice are associated

				well as the rate of missing data, were not examined.		with increased response set.
Mavaddat et al, 2009 (81)	✓✓ The PEQ instrument appears to be a valid and reliable instrument, able to assess patients' views of the quality of primary care mental healthcare at practice level. Although there were more focus groups with SMI patients than CMHPs patients due to different recruitment methods, there were no differences between groups' views of primary care mental health.	✓ Face validity was ensured by the item development process, which relied on focus groups with patients. Items were also reviewed by members of the National Institute of Mental Health England (NIMHE) primary care service users steering group, practice staff and GPs to ensure that the questionnaire was also acceptable to healthcare professionals.	✓✓The items' content was established from the patients' perspective using separate focus groups for people with severe and persistent mental disorders and those with common mental disorders.	✓✓ Descriptive statistics for each item was computed to evaluate the distribution of scores. All items had responses that included the full range of the Likert scale. The initial 30-item version, applicable to all respondents, was reduced to improve the usefulness of the questionnaire. This shorter version was based on alpha reliability scores after eliminating items with the lowest item-total correlations.	✓ All the items included in the final 20-item version showed factor loadings greater than 0.30 although some had cross-loading on both factors. Cronbach's alpha was high (0.94).	✓✓
Mayston et al, 2017 (82)	✓✓ The MHSSS instrument is a useful measure of satisfaction with services, but it	✓ Face validity was partially ensured by the item generation process that included the views of service	✓✓ The initial measure was developed from the perspective of service users and their caregivers through	✓✓	✓ For a 1-factor solution, all factor loadings were acceptable (>0.40), except for 4 items for which they were nevertheless greater	✓✓

	would be informative to investigate how the final MHSSS performs in other independent samples and other comparable settings.	users and their caregivers. However, service users were not included as part of the expert panel that reviewed the draft instrument.	individual interviews and focus groups. Participants were also asked to comment on existing measures, whose items were revised in light of the qualitative findings. Additional items were added after the draft was reviewed by an expert panel, to cover relevant aspects of the study setting that were not mentioned elsewhere.		than 0.30. Cronbach's alpha was high (0.92).	
McGuire et al, 2007 (83)	✓✓ Although the instrument was necessarily developed and tested within a selective sample, STAR is a brief scale with good psychometric properties to assess therapeutic relationships in community psychiatry.	✓ Face validity was ensured to some extent by the instrument development process which involved patients at all stages. But the STAR-P instrument was not subjected to a pre-test phase before being administered on a larger scale.	✓✓ Two approaches were used to generate a first set of items: semi-structured interviews were conducted with clinicians and patients and a literature review. To do this, 9 existing scales were identified and submitted to participants to evaluate their applicability to community care setting.	✓✓ A number of items were removed based on the results of factor analyses. Subsequently, other items with the lowest test-retest reliability on each factor were dropped to develop a brief scale. Missing data as well as ceiling and floor effects were not examined.	✓✓ All items had a factor loading greater than 0.50 and Cronbach's alpha was >0.70 in all domains.	✓✓
Meehan et al, 2002	✓ Although the	✓✓ Face validity	✓✓ Items were	✓✓ Dispersion of	✓✓ All items had factor	✓✓ A 5-response

(84)	IESQ instrument was developed in the acute care setting, it is a brief, user-friendly instrument that seems to perform equally well in the rehabilitation setting.	was ensured by the item development process that involved patients. First, inpatients (n=66) were asked to identify aspects related to their satisfaction with their hospital stay. Second, another sample of inpatients (n=72) rated these items in terms of their importance to reduce the initial pool.	developed from a literature review combined with the patients' perspective.	responses was good and the IESQ instrument had high internal consistency. Therefore, all items were retained at the end of the analyses.	loading greater than 0.40. Cronbach's alpha was >0.70 for the total scale and in all domains.	format was used because it produces good response variability across all response options.
Misdrahi et al, 2009 (85)	✓ The 4PAS instrument was designed to be administered in any clinical setting for schizophrenic patients, but precautions must be taken because it was only tested in a population of hospitalized patients.	X	X Patients were not involved in the development process of the 4PAS instrument.	✓✓ All items submitted for psychometric evaluation were retained at the end of the tests. Except for one, the set of items demonstrated solid properties. However, this item was kept because its deletion did not substantially modify the average inter-item correlation or the	✓ All items had a factor loading >0.40 except for one item that had a cross-loading on both factors. This item was assigned to the factor with which it shared the closest conceptual relationship. Cronbach’s alpha was greater than 0.70.	✓✓

				Cronbach alpha. Further analysis is required to determine whether this item should be deleted or rewritten.		
Moutoussis et al, 2000 (86)	✓ The survey questionnaire items were derived from an existing validated instrument that was modified for the purposes of the study. Caution should be taken as the resulting questionnaire was not subjected to extensive psychometric analysis prior to use.	✓✓A users' group of the CHC reviewed draft versions of the questionnaire to assess its acceptability and appropriateness. The group was broadly satisfied with the measure.	✓✓ The instrument was adapted from an existing questionnaire developed for inpatient psychiatric patients. It was tailored to the outpatient care context by identifying specific aspects from the literature and input from users.	X	X	✓✓ A five-point Likert scale was used to offer a wide range of possible responses.
Nabati et al, 1998 (87)	✓ The SI-MH instrument is a brief, easy-to-use instrument with reasonable psychometric properties in ambulatory mental health patients.	X	X The questionnaire was developed from an existing instrument, the Satisfaction Index, which was widely used in primary care setting. The original tool was adapted by simply inserting the word	✓✓ Ceiling and floor effects, as well as the rate of missing data, were not examined. The exploration of the psychometric properties of this instrument in a	✓✓ All items had a factor loading greater than 0.40 and Cronbach's alpha for the total scale was high (0.90).	✓✓

	However, sensitivity to change was determined with a subsample of bipolar patients.		“mental” before “health care”. Patients were not involved at any time in the instrument development process.	mental health clinical sample demonstrated reasonable properties and all items were kept.		
Nordon et al, 2013 (88)	✓ The instrument was designed to assess satisfaction with psychotropic medications in patients regardless of their disorder, but the psychometric properties were assessed only in bipolar patients with manic or mixed episode. Further studies should replicate these findings to other psychiatric disorders.	✓✓ Individual semi-structured interviews with psychiatric outpatients (n=30) were conducted to ensure that scale captured the important aspects of psychotropic treatment, to assess and optimize item formulation and to ensure a good understanding of each item.	✓ The items were generated from an extensive literature review. Patients were asked to participate in the subsequent item reduction phase through semi-structured interviews.	✓✓ A first phase of item reduction through semi-structured interviews with patients was carried out. Non-response rates to each item was low (<0.4%) and there was no floor or ceiling effect. Two pairs of items had high inter-item correlations, but they were kept because they seemed to provide specific information. Moreover, one item had modest correlations with the others and its factor loading was less than 0.40. This item must be	✓✓ Factor loadings were all greater than 0.4, except for one item (0.38) and Cronbach's alpha was good (0.85).	✓✓

				reformulated and validated in a further study.		
Oades et al, 2011 (42)	✓✓ The CEO-MHS instrument appeared to be useful with inpatients and outpatients.	✓✓ The involvement of consumers, as researchers and users, in the item generation process was ensured the face validity. The items were then reviewed by subgroups of consumer researchers, university-based researchers, mental health experts and consumers, leading to some refinements. In addition, a small pilot test was carried out with consumers (n=7) to ensure that the items were understandable.	✓✓ Consumers were interviewed through focus groups and individual interviews to identify all important aspects of mental health service evaluation from their own perspective. The items were then generated using a brainstorming method by a research team of consumer researchers.	✓✓ First, the initial pool of items was reduced to balance the requirements of reliability with ease-of-use. A second phase of item reduction was undertaken using a factor analysis.	✓✓ All items included in the final version had high factor loadings (>0.50) on a unique factor. Cronbach's alpha was high for total scale (0.92) and in both domains.	✓✓ A 5-point Likert scale was adopted because it can approximate a continuous scale more easily, while providing information on the direction and intensity of responses.
Rose et al, 2011 (89)	✓✓ The CEO-MHS instrument is a useful measure of consumer satisfaction with mental health services and thus an important tool for the development of person-centered mental health services organization and delivery.	✓ Given the potential usefulness of the original questionnaire, its applicability to NGO mental health services in Australia was explored. Subsequent modifications were	✓✓ Two items were removed from the initial questionnaire because they did not apply to NGO mental health services in Australia, and then two additional items on rights and responsibilities were added following suggestions from a consumer advisory	✓✓ A reduced set of items was obtained from an automatic backward stepwise regression analysis. To achieve this, items that were significantly related to the question of consumer satisfaction were then matched to one item that was most representative	✓ Cronbach's alpha was high (0.84) for this reduced set of items. Additionally, it was suggested that the reduced data set yields a better fit to the data and that the single factor model is the more robust solution (but factor loadings not shown).	✓✓

		made according to the suggestions of a consumer advisory committee.	committee.	for each theme in the original instrument (based on the strongest standardized beta weight).		
Ortiz et al, 2012 (90)	✓✓ The overall findings provide evidence that the ICS instrument is a reliable measure of consumer satisfaction in psychiatric inpatient settings.	✓✓ The initial 43-item version was tested in a pilot study with consumers (n=1 027) and reduced to 28 items after factor analyses. This adapted version was then carefully reviewed by a workgroup of consumers and professionals who commented on the wording, readability, length, redundancy, and clarity of each item.	✓✓ This study consisted of formulating an inpatient version of the MHSIP Consumer Survey. The instrument development process involved a number of stakeholders, including both patients and staff through a series of meetings.	✓✓ The revised 28-item version was analyzed in terms of skewness, kurtosis, missing value and inter-item correlations and no items were removed according to these criteria. As is usually the case with satisfaction data, all items were positively skewed. Floor and ceiling effects ranged from 2.0% to 5.4% and 4.0% to 16.4% respectively. The average rate of missing data was low (2.6%). Seven items were excluded from the psychometric analyses because they did not relate	✓ All items with factor loadings <0.30 or not loading in any domain were dropped from the instrument (data not shown). Cronbach's alpha was > 0.70 for the ICS total scale and in all domains.	✓✓

				to any of the identified domains, but were retained in the final version of the instrument due to their relevance for users.		
Parker et al, 1996 (91)	✓✓ This is an initial study to describe the first stage of development of a patient satisfaction questionnaire that addresses domains relevant to private and public outpatient psychiatric practice. Nearly all sample members were attending private practitioners, which may explain the low ratings on some items.	X	X	✓ The initial 62-item form was reduced to 43 items using a factor analysis, by deleting all items with a loading less than 0.30 and those with a cross-loading. Ceiling and floor effects, and missing data, were not examined.	✓✓ All items retained in the final version had a factor loading equal to or greater than 0.30 on a single factor. Cronbach's alpha was not computed.	✓✓
Pellegrin et al, 2001 (92)	✓ The CPOSS instrument was designed for use with adult psychiatric outpatients, but was only tested with a	X	✓ A number of items were selected from an inpatient satisfaction scale, while other items were developed for the outpatient setting based on the input from	✓ Following the evaluation of the psychometric properties of the CPOSS instrument, none of the 15 initial items were	✓ Cronbach's alpha for the total scale was >0.70, but no factor analysis was performed.	✓✓ A 5-point Likert scale was used to minimize positive response bias and optimize variability and predictive validity. In addition, a

	sample of outpatients from clinics affiliated with a public-academic psychiatric institution. This scale should be further tested to ensure its generalizability to psychiatric outpatients in other settings.		psychiatric clinicians and a comprehensive review of studies that used surveys and focus groups to identify service features that are important to psychiatric patients.	dropped. However, ceiling and floor effects, as well as the rate of missing data, were not examined.		4-point response format was used to assess behavioral intent to recommend the clinic.
Perreault et al, 2001 (93)	✓✓ Both original and adapted versions had strong psychometric properties, so the 21-item adapted version seems justified for use as a general satisfaction scale.	✓Face validity was partially ensured by the instrument generation process that included patients through qualitative interviews.	✓ The 35-item original version was developed based on a literature review and the view of administrators and staff from outpatient clinics. The reduced version obtained from a factor analysis was then enhanced with two items from patients' comments gathered from interviews on the original version.	✓The distribution of results showed very little variance and a strong halo effect, however, rate of missing data was not addressed.	✓✓ Factor loadings were >0.40 for all items included in the final 21-item version. Cronbach's alpha was >0.70 for the total scale but two of the five domains had a coefficient below the cut-off.	✓✓ Use of a 4-point scale alternately reversed to minimize the "halo" effect on the distribution of responses.
Perreault et al, 2006 (94)	✓✓ The PPIQ instrument appeared to be helpful in measuring client	✓	✓✓ The instrument development process used a combined approach that included a	✓ Response distributions were examined for normality through	✓ Cronbach's alpha >0.70 for the two PPIQ subscales. No factor analysis was	✓✓

	perspectives and satisfaction with different components of information in a mental health outpatient setting.		literature review and the perspective of patients and professionals.	calculations of skewness and kurtosis. For both subscale, scores were within the normal range. The PPIQ instrument demonstrated strong psychometric properties and as a result, all items subjected to analyses were kept in the final version.	performed.	
Peytremann-Bridevaux et al, 2006 (95)	X	X	X	X	X	✓
Phattharayuttawat et al, 2005 (96)	✓✓ The TPSS instrument is a measure of patients' satisfaction designed specifically for outpatient psychiatric clinics. This scale has good psychometric properties but should be tested further to confirm its advantages and	✓✓ The instrument was reviewed and refined on several occasions by experts. A pilot study was conducted prior to large-scale administration and modifications were made accordingly.	X The items were constructed from a comprehensive review of previous studies that have identified areas important to psychiatric patients. The content of these instruments was then discussed in focus groups with experts.	✓ Distribution of responses was not analyzed in terms of missing data and ceiling or floor effects. All items loading < 0.4 on any factor were removed from the scale.	✓✓All items loading >0.40 on each factor constituted a separate scale of the TPSS instrument. The scale had an excellent overall Cronbach's alpha coefficient (0.96), as well as all dimensions (>0.70).	✓✓

	ensure its generalizability to psychiatric outpatients in other settings.					
Priebe et al, 1993 (97)	X	X	X	X	X	✓✓ Four items used a 100-mm VAS that can be interpreted as an 11-point rating scale and one item had a dichotomous response format.
Rofail et al, 2005 (98)	✓ The aim of this study was to develop a scale to measure patient satisfaction with antipsychotic medication, however the sample was restricted to people with schizophrenia.	✓✓ The preliminary version was subjected to a sample of patients (n=10) to assess acceptability, clarity, and understanding of each item. According to these feedback, a second draft questionnaire was produced and submitted to a pilot study (n=69). Finally, the resulting version was then scrutinized by potential users and submitted to a cross-section of patients in	✓✓ The SWAM questionnaire was developed from the perspective of patients and professionals and a literature review. Furthermore, in the preliminary version, spaces have been provided for comments to identify any other important dimensions.	✓ Firstly, some items were eliminated as a result of patient reports in pilot study. Data sample included only one case with missing data. Score distribution was not analyzed in terms of ceiling or floor effects. According to the results of the factor analysis, a number of items were discarded because they did not load on either factor or they were	✓✓ All items included in the final 23-items version had a factor loading > 0.40. Cronbach's alpha was 0.91 for the total scale and was >0.70 in all domains.	✓✓ A 5-point Likert scale was used to provide a reasonable range of response options, especially given the cognitive dysfunction associated with psychosis.

		a preliminary study (n=10) and the questionnaire was modified accordingly. Additionally, an advisory team and a panel of professionals attested to the relevance and usefulness of the measure.		not conceptually representative of the factor to which they were intended to be related.		
Rose et al, 2009 (99)	✓ The CONTINU-UM instrument was developed and validated for a narrow patient group recruited from community mental health teams.	✓✓ The draft measure was reviewed and refined by two expert groups (n=6 and 4), themselves made up of service users. In addition, the draft version was subjected to a small consultation exercise with professional experts (n=2) and one user researcher. Finally, a feasibility study was carried out with service users (n=37) to ensure that the measure was	✓✓ The instrument was entirely generated from the perspective of service users using focus groups.	✓ The items related to the choice domain were removed due to a poor conceptual fit, as were the items in the importance dimension because they had a skewed distribution of scores, resulting in a lack of variance. Also, there is very little missing data, but substantial not applicable responses in some domains.	X	✓✓

		understandable and easy to complete. Minor changes have been made as a result.				
Røssberg et al, 2003 (100) **Røssberg et al, 2003 (101)**	✓✓ The revised subscales had acceptable psychometrics andgave a clearer picture of the relationship between the perceived level ofpatient satisfaction and the WAS subscale scores.	X No evidence was provided regarding the evaluation of the face validity of the revised version of the WAS.	X Some modifications were made to the original instrument due to the need to modernize the WAS as a result of cultural changes leading to weakened psychometric properties. Thus, some items have been replaced, and others added, from an existing scale, the Community-Oriented Programs Environment Scale. In addition, an additional clinically significant dimension was identified and conceptualized within a new subscale.	✓ First, some items were deleted from the original WAS instrument because they were considered outdated and no longer clinically acceptable, resulting in a variance close to zero for these items. Other items were then removed because they had low internal consistency and low item-to-subscale correlations. Furthermore, their removal has improved Cronbach's alpha for each dimension.	X The factor structure of the WAS is difficult to establish as described previously. The strong intercorrelations among several subscales assume a structure with less than 10 dimensions. However, no factor analysis was performed for this revised scale. Although improved from the original version, Cronbach's alpha was less than 0.70 for the total scale and in nine of the eleven subscales.	✓✓ To enable patients to nuance their ratings, the initial true/false format was transformed into a 4-point response scale.
Ruggeri et al, 2000 (102)	✓ The VSSS-EU is an instrument for	✓✓ The original questionnaire was	✓ The original Italian version VSSS-54 was	✓ All the items initially developed	✓ Cronbach alpha was greater than 0.70 for the	✓✓ The majority of the items used a 5-

Ruggeri et al, 1993 (124)	measuring patient satisfaction with mental health services across Europe. It showed good overall psychometric properties, but it was only used in a sample of people with schizophrenia.	pilot-tested with professionals (n=5) and service users (n=10) and some changes were made accordingly. Subsequently, respondents were also asked to identify important aspects not covered by the questionnaire. Finally, the five language versions of VSSS-EU used a focus group methodology to culturally adapt the underlying instrument concepts and all of these versions appeared acceptable.	developed partially from a literature review. Almost one-half of the items were adopted from existing scales covering aspects intended to be relevant in a wide range of medical and psychiatric contexts. The other half was developed by the authors to be relevant to the psychiatric and Italian care settings. VSSS-EU is the result of a cross-cultural validation and was produced in five European languages (Danish, Dutch, English, Italian and Spanish). Specific items were changed to adapt the instruments to the context of each country’s mental health system.	were retained after the psychometric tests.	pooled sample and in the different EPSILON sites. By contrast, alpha did not perform well for three of the seven dimensions in various samples, particularly the access dimension, which had a value of 0.06 for the pooled sample. No factor analysis was performed.	point scale. The other items used a yes/no response format for screening questions. When the answer was yes, participants answered with the same 5 point scale while when the answer was no, they answered with a 3 point scale. These formats made it possible to estimate the subject's degree of satisfaction both with the interventions provided and with the professional's decision not to provide an intervention when this was the case.
Rush et al, 2013 (103)	✓✓ The OPOC-MHA instrument is a valid and reliable perception of care tool for clients and their family	✓ Qualitative data about the instrument was collected only from the staff perspective, through feedback from pilot	X The list of items was developed from a comprehensive literature review and an environment scan to determine what	✓✓ The rate of missing data was generally low (less than 4.6%), while the "non-applicable"	✓ Several exploratory factorial analyses were performed according to the nature of the correlation matrix. All analyses resulted in a 4-factor structure and	✓✓

	members/supporters receiving services within mental health and addiction agencies.	site contacts, phone interviews and an online survey with site leads and staff. These feedbacks indicated that the OPOC-MHA was generally well accepted and comprehensive. Only minor revisions were made to the structure and language of the instrument.	measurement tools and processes were already in place.	response rate was higher for some items which pertained to a small number of the respondents. Data were skewed due to a ceiling effect. However, no items were removed based on the results of descriptive or factor analysis.	the same pattern in items loadings. The resulting factorial solution was consistent with the preliminary conceptual structure. All items had a factor loading of at least 0.40. A few items had high cross-loadings and were subsequently assigned to their conceptual domain. Cronbach's alpha was >0.70 in all domains.	
Schalast et al, 2008 (104)	✓✓ The findings showed that the EssenCES questionnaire is a valid instrument for assessing the ward atmosphere in forensic psychiatry.	X	X Lists of possible items were drawn from a series of three studies using scale and factor analyses to select appropriate items.	✓ Ceiling and floor effects and the rate of missing data were not examined. But all items submitted for psychometric analysis were kept.	✓✓ All items reached a factor loading of at least 0.40. Cronbach's alpha was >0.70 in all domains.	✓✓
Schröder et al, 2007 (105) Schröder et al, 2010 (106)	✓✓ The QPC-IP instrument is psychometrically adequate for evaluating patients' experiences of the quality of psychiatric care.	✓✓ Firstly, the initial pool of items was reduced by half after being reviewed by the authors three times. Then, face validity was tested in a pilot study with a sample of	✓ Items were drawn from the results of an earlier qualitative study conducted with a sample of in- and out-patients (n=20).	✓✓ The initial pool of items was successively reduced based on the authors' review of the measure, followed by the findings of a pilot study with a sample of patients. The	✓✓ All items in the final version were equal to or larger than 0.40 and none had cross-loading. Cronbach's alpha was satisfactory for the total scale and in all domains.	✓✓ A 4-point Likert scale was used to avoid neutral answers. The same response scale was used in the in- and out-patient versions.

Schröder et al, 2011 (107)	✓✓ The QPC-OP instrument includes important aspects of patients' assessments of quality of care and shows adequate psychometric properties.	experienced in- and out-patients (n=6). The questionnaire evaluation was carried out in several stages: first, patients were asked to complete the draft and evaluate it in writing. Later, patients discussed with researchers about the questionnaire (face-to-face or by phone). Clarity and comprehensibility, as well as importance of each item, were assessed and debated into group discussions. Consequently, a number of items were deleted for several reasons: item's content appeared less important or endorsed by few participants, or also because they were viewed as emotionally charged	resulting 69-item version was used as the basis for the development of the in- and out-patient versions. Items were then deleted from the original version following step-by-step factor analyses for the following reasons: insufficient sampling adequacy (<0.60), factor loading less than 0.40 or cross-loading on another factor. ✓✓ Two items were excluded because they did not load sufficiently large on either factor.	✓✓ All items had a factor loading >0.50 on their respective factors. Cronbach's alpha was high for the total scale (0.95), but below the recommended cut-off for two of the eight domains.

		or overlapping.				
Shiva et al, 2009 (108)	✓ The study sample was restricted to male because forensic inpatient psychiatric services provide mental health care only to male inmates.	✓ The initial pool of items was reviewed by a panel of professionals. They were asked to examine the items keeping in mind the patients' general level of education and appropriateness to setting. However, patients were not included in any step of the instrument development process.	X The survey items were drawn from a review of existing measures and then approved by consensus among the forensic staff members and psychiatry administration.	✓✓ Two items were dropped because they were not significantly correlated with the total scale score and did not adequately differentiate between low and high total scores. Other items were dropped because their factor loading did not meet the 0.40 criterion, or they were loaded on more than one factor. Finally, items were deleted during the CFA when parameter estimates were non-significant, SMCs were low or the standardized residuals were 2.58 and higher.	✓✓ All item loadings were greater than 0.40 and Cronbach's alpha was >0.70 for both forms of the ISQ and their respective factors.	✓✓
Slade et al, 2014 (109)	✓✓ The CDIS instrument is the first short, standardized	✓✓ The draft version in the original language was evaluated in two	✓✓ The instrument was developed from a literature review combined with the	✓ Distribution across the range of scores was investigated in	✓ Cronbach's alpha was >0.70 for the Satisfaction subscale. No factor analysis	✓✓ The CDIS instrument initially used a 5-point Likert scale for both

	measure of involvement and satisfaction with a specific decision related to mental health care, which is suitable for use across a range of clinical settings and available in five languages.	ways. First, it was evaluated by focus groups with service users (n=7) and staff (n=7). After refinement, the draft was then piloting with service users (n=9) and staff (n=7) to assess its feasibility. In addition, all target language CDIS were validated through focus groups (n = 38) and piloting with a sample of service users (n=30) and staff (n=31) and then submitted to a cognitive debriefing in each country. No modifications were required. The final versions were then approved at a study meeting.	views of service users and staff to ensure coverage of all relevant domains.	terms of ceiling and floor effects. The Involvement sub-scale had appropriate distribution and no floor or ceiling effects, while the Satisfaction sub-scale was right-skewed as is typical with satisfaction data. Additionally, the latter was analyzed both as continuous and categorical after collapsing the response categories according to clinical meaningfulness. All initial items were kept after psychometric evaluation and the CDIS scales demonstrated adequate properties.	was performed.	subscales. However, for the Involvement subscale, the response categories were collapsed into 3 modalities because they reflected a social desirability bias rather than different experiences.
Slater et al, 1982 (110)	✓ The SMHC scale was designed to measure satisfaction	✓✓	✓ The items have been developed exclusively from the professionals'	✓✓ During the analysis of the instrument	✓ All items had a factor loading greater than 0.40, except for four items that	✓✓ A 4-point scale was chosen to force either a positive or

	with outpatient mental health care and it would appear that it is also useful for assessing inpatient treatment, but this was not tested.		viewpoint. Patients were only asked to evaluate their understanding of the items during a pretesting. Some items were eliminated during this step.	properties, a number of items were discarded for several reasons: the findings of the pre-test phase, low variance, ICCs less than 0.50, very low loading on factors or because they constituted a separate and isolated factor.	were included because they seemed relevant to the dimension. Cronbach's alpha was not computed.	negative response rather than providing a midpoint value of uncertain.
Speckens et al, 2000 (111)	✓ The RQ instrument was demonstrated to have psychometrically sound properties and appeared to be a useful instrument to assess reassurability in medical patients and to evaluate the effect of interventions by physicians or psychotherapists. Further research questions concern whether the factorial structure of the present study	X	X	✓✓ Two items were removed because they did not have sufficient factor loadings to the common factor for all three samples. Ceiling or floor effects, as well as the rate of missing data, were not examined.	✓ Items had factor loadings greater than 0.40 for all three samples, with the exception of one item (0.298) in population 2, but it was still retained in the final version. Cronbach's alpha was overall satisfactory, although it was slightly lower for population 1 (0.66).	✓✓

	can be replicated in other patient populations.					
Svedberg et al, 2008 (113) **Svedberg et al, 2007 (112)**	✓ One of the limitations of the study is the sample size which can have an impact on the generalization of the results. It is therefore stated that the study findings can only be considered as representative of the study participants.	✓✓ The draft questionnaire was first discussed in the group of co-authors to secure the meaning and wording of the items. Secondly, the questionnaire was pilot-tested with a sample of outpatients (n=15) to check whether the items communicated the intended meaning, as well as the feasibility and usefulness of the HPIQ instrument. Some modifications were made according to this feedback.	✓ Items were generated using patients' and nurses' perspectives gathered from two previous qualitative studies.	✓✓ Several items were removed during the pilot study to reduce the risk of misinterpretation and the rate of missing data. Secondly, eleven items were excluded due to too many non-responses, a high cross-loading or a factor loading lower than 0.45.	✓✓ All factor loadings in the final version were higher than 0.50 and Cronbach's alpha was 0.90 for the total scale and greater than 0.70 in all domains.	✓✓
Uijen et al, 2011 (114) **Uijen et al, 2012 (115)**	✓ The NCQ instrument is intended to be used to compare continuity experiences for	✓✓ Face validity was tested on a sample of 15 patients with chronic diseases using the "thinking aloud	✓ Items were drawn from existing instruments and additional items were derived from qualitative interviews with patients	✓✓ Before perform factor analyses, some items were excluded because they were not sufficiently	✓✓All items had a factor loading greater than 0.40 and Cronbach's alpha was high for all domains.	✓✓

different diseases and multi-morbidity patterns. Despite potential recruitment bias, patients were included from both primary and secondary care.	technique". Changes were made accordingly and tested in subsequent interviews.	(n=30) conducted in a previous study.	distinctive for measuring continuity of care or had poor variability. Data distribution was analyzed in terms of ceiling and floor effects and missing data. The negatively worded items were deleted because they did not perform well during PCA. The factors were then refined by excluding some items for the following reasons: high inter-item correlations, high ceiling effect and low SD, high cross-loading or relative low loading on all factors, high missing rate and decreases the interpretability of the factor. The deletion of these items resulted in an improvement of the

				internal consistency of the scale. The final version did not have a floor or ceiling effect.		
Ul-Haq, 2012 (116)	✓ The survey questionnaire items were derived from an existing validated instrument. Caution should be taken as the psychometric properties of the instrument were not documented.	X	X	X	X	✓✓
Ware et al, 2003 (117)	✓ The results showed that the measure was easy to administer and produced well-distributed responses among persons with serious mental illness. However, this field test was conducted only in the greater Boston area limits its generalizability. A second psychometric	✓✓ Firstly, the draft versions were subjected to cognitive interviews (n=9) to examine the meaning and relevance of items and to assess the feasibility. Then, two pilot studies were carried out with 41 individuals with severe mental illness, resulting in some revisions. Following this	✓ CONNECT instrument was developed using data from a previous ethnographic study and clinical experience. The conceptual domains composing CONNECT are in line with those found in the literature.	✓ Data distributions were examined during pilot tests and some modifications were made accordingly. Items were added to improve internal-consistency reliability and some skewed items were re-written to reduce ceiling effects and improve score distributions.	✓ Cronbach’s alpha was >0.70 in all subdomains. No factor analysis was performed.	✓✓ Five-point Likert scales were used with different response categories to match the items' content. Respondents reported no difficulties with this scoring method.

	evaluation should include a more focused validation effort informed by these results.	phase, a number of items were rewritten or deleted. Most respondents seemed to understand the items well and emphasized the usefulness of the measure.		Additionally, several items were deleted to improve scaling success rates and others were revised for a better fit with item stems. During the field test, floor and ceiling effects, as well as missing data rates, were low. All of the items submitted for field testing demonstrated robust psychometric properties and were retained.		
Webster et al, 2012 (118)	✓ Although the study used a pre-validated scale to partially ensure content and face validity, caution should be taken as the resulting questionnaire was not subjected to extensive psychometric analyses prior to	✓This instrument used as a starting point a pre-validated questionnaire that was improved based on results of nurse interviews and patient focus groups, that addressed all variants of satisfaction that are relevant to anxiety management. The	✓✓ The items were adopted from an existing questionnaire developed by the same health care department where the current study is conducted. This questionnaire was developed based on literature review, in-depth interviews of hospital staff and focus groups of patients and	X	X	✓✓ Use of the same response scale as the original questionnaire

	use.	additional items were then submitted to a panel of experts to assess their appropriateness; however, this panel did not include service users.	nursing staff. The questionnaire was supplemented by additional items specifically generated for the purpose of this study.			
Wongpakaran et al, 2013 (119)	✓ Patients were recruited as part of a group therapy whose enrollment depended on the clinician's judgment, so selection bias may have occurred.	X	X Items were derived from existing scales.	✓✓ Before being subjected to psychometric tests, distribution of response data was examined. There were low rate of missing data and distribution was broadly normally distributed (skewness and kurtosis coefficients <3). The significant outliers were eliminated. All items were then submitted to factor analyses and were kept in the final version.	✓✓ All items had a high factor loading ranging from 0.52 to 0.83 and Cronbach's alpha was good (0.87).	✓✓
Woodring et al,	✓✓ The PSIPSS instrument	✓✓ The initial version of the	X No details were provided about the	✓ Five items were deleted from the	✓ All items had a factor loading greater than 0.40.	✓✓

2004 (120)	demonstrated acceptable reliability and validity in the study sample, and therefore could be applied to assess patient satisfaction in other healthcare setting.	PSIPSS instrument was pilot tested with a sample of patients (n=30) to assess the appropriateness, ease of administration, and interpretability of items. Items were continually reviewed by experts in psychiatry care and outcomes research to establish face and content validity. Also, a substantial number of patients (n=182) indicated that some aspects were not covered by the questionnaire statements, but most of these comments were highly specific to the individual's hospital experience.	development process of the initial 20-item version, only the item selection process was described.	initial version based on the results of the pilot test, but the selection process was not clearly detailed. The score distribution was positively skewed and the rate of missing data was not examined, but no examination of ceiling or floor effects, as well as rate of missing data. The 15 items submitted to the psychometric tests were all kept in the final version.	For items that had cross-factor loadings, items were assigned to a single factor based on their conceptual relationship with the other items. The internal consistency reliability was high for both the total scale and in all domains.	
Wright et al, 2006 (125)	✓✓ The SHEP survey provided a useful tool for evaluating and improving satisfaction among VHA veteran users. Because of the	X	✓ The dimensions were derived from instruments developed by the Picker Institute and refinements to these instruments were made in light of veterans' experiences through	X	✓ Cronbach's alpha was below the recommended threshold in three dimensions for both the inpatient and outpatient surveys.	✓

	disproportionate low number of women served by the VHA, these findings may not generalizable to non-veteran populations.		nation-wide focus groups of veteran patients and their families conducted in a previous study.			
Zendjidjian et al, 2015 (121)	✓ The representativeness of the sample is one of the limitations of the study. Patients were recruited from two psychiatric hospitals; therefore, the psychometric properties need to be studied in other type of hospitals and other countries.	✓✓ Patients were asked to comment on any aspects of the questionnaire and no items were found irrelevant or in need of improvement. The SATISPSY-22 proved to be easy to use.	✓✓ Items were generated from the patients' exclusive perspective and reflected concerns and perceptions that were of great importance to them.	✓✓ A number of items were removed due to low response rate (>10%), low index discrimination (<0.80), high inter-item correlation (>0.80), and after examination of items' meaning and structure. Floor (2.2 to 13.8%) and ceiling effects (18.0 to 40.1%), as well as the rate of missing data, were satisfactory. The remaining items showed INFIT statistics inside the acceptable range. SATISPSY-22 instrument demonstrated	✓ All items had a factor loading greater than 0.40 and those with a cross-loading were kept and assigned to the factor with which they had a closer conceptual relationship. The unidimensionality of each dimension was assessed using a Rasch analysis. Cronbach's alpha was >0.70 in all dimensions.	✓✓ The range of response options was determined according to the patient interviews.

				strong psychometric properties.		
Zimmerman et al, 2017 (122)	✓ The demographic profile of study participants could not be described because patients completed the CUPSS anonymously, limiting any conclusions about the representativeness of the sample. The study should therefore be replicated in samples with different demographic and clinical characteristics. Additionally, the generalizability of any single site study is somewhat limited.	✓✓ The initial list of items was reviewed by clinicians and members of the Department's Quality and Outcome Committee. In addition, response burden to fill the scale was examined and most patients (93.1%) considered it as a minimal or not at all burdensome. An open-ended question was added to obtain any additional comments. Patients were also interviewed in a pilot study to suggest items to include on the scale.	✓ Items were generated from the clinical experience and a literature review that included patient focus groups studies.	✓ Descriptive statistics were computed for each item. Similar to other satisfaction scales, all items were negatively skewed. Only one item was eliminated because it was highly correlated with another item (>0.75). All items met the expected standards, demonstrating strong psychometric properties of the CUPSS instrument.	✓ Cronbach's alpha was >0.70 for both samples. No factor analysis was performed.	✓✓ A 5-point Likert scale was used to reduce skew and increase variability of responses and thus increase the likelihood of detecting differences between clinicians.

Abbreviations: NA, not available

✓✓ - positive rating, ✓- acceptable rating, X - negative rating.

Table 4. Evaluation of the instruments' performance.

Reference	Convergent validity	Discriminant validity	Predictive validity	Test-retest reliability	Responsiveness
Aloba et al, 2014 (43) **Anderson et, 1990 (44)**	✓ There were weak significant positive correlations between TPS score and Medication Adherence Questionnaire score ($r=0.257$, $p<0.001$) to support construct validity.	X	X	X	X
Atkinson et al, 2004 (45)	X	X	✓ Preliminary evidence support the predictive validity of the TSQM scales as a good predictor of patients' medication adherence. Thus, regression analyses showed that in six of the eight illness groups, 50-60% of the variation patients' ratings of their	X	X

			likelihood to persist with their current medication was predicted by TSQM scores. However, for two illness groups, this result was more contrasted (respectively, 42% and 26% for asthma and hypertension, respectively).		
Baker, 1990 (46)	X	X	X	X	X
Barker et al, 1999 (47) **Barker et al, 1996 (48)**	✓✓ By using Spearman's correlation coefficient, the PCSQ instrument had correlations of 0.72 with the Client Satisfaction Questionnaire (CSQ) and 0.82 with the Scale for the Measurement of Satisfaction with Medical Care (SMSMC).	X	X	✓ Test-retest reliability was evaluated one week after the initial administration. Cohen's kappa coefficient ranged from 0.48 to 0.80.	X
Berghofer et al, 2011 (49)	✓ Convergent validity was established based on significant correlations between the ECS factor scales 1, 2 and 3 and both factors of MHSIP-Consumer Service Satisfaction Survey (Pearson's coefficient r ranging from 0.26	X	X	X	X

	to 0.46, p<0.01). However, the ECS factor scale 4 had a weak (or no) correlation with the MHSIP factors because the two instruments measure different aspects of treatment satisfaction. In addition, there were significant correlations between the overall satisfaction item and the ECS total scale (r=0.72, p<0.01) and with the four ECS factor scales (r ranged from 0.32 to 0.68, p<0.01).				
Bjertnaes et al, 2015 (123)	X	X	X	✓✓ A retest questionnaire was administered approximately 2 days after the first completion. The intraclass correlation coefficients (ICCs) ranged from 0.83 to 0.84 and supported the test-retest reliability of the PIPEQ-OS version.	X
Blais et al, 2002 (50)	X	X	X	X	X
Bramesfeld et al, 2007 (52)	X	X	X	X	X
Brunero et al, 2009 (51)	X	X	X	X	X

Bruyneel et al, 2018 (39)	X	X	X	X	X
Caruso et al, 2013 (53)	X	X	X	X	X
Clement et al, 2012 (16)	✓✓ Convergent validity for the BACE treatment stigma subscale was supported by a significant positive correlation with the Stigma Scale for Receiving Psychological Help ($r = 0.30$, $p = 0.001$) and the Internalized Stigma of Mental Illness scale ($r = 0.40$, $p<0.001$).	X	X	✓ Test-retest reliability was assessed in two ways. Lin's concordance statistic for the treatment stigma subscale was 0.82, indicating an acceptable reliability. Kappa's coefficient was >0.40 for all but one item (0.35) and ranged from 0.61 to 0.80 for the majority.	X
Eisen et al, 2001 (54)	X	X	X	X	X
Eisen et al, 2001 (54)	X	X	X	X	X
Eisen et al, 1999 (55)					
Eisen et al, 2002 (56)	X	X	X	X	X
Eton et al, 2017 (58) **Eton et al, 2015 (57)**	✓ Convergent validity was supported by significant correlations between all PETS domain scores with those of conceptually-related constructs assessed in established	X	X	X	X

	measures (all significant at p<0.001). Higher PETS scores were associated with Chronic Condition Distress scale (r ranged from 0.40 to 0.65), TSQM side effects (r ranged from -0.22 to -0.39), TSQM convenience (r ranged from -0.28 to -0.51), PMCSM (r ranged from -0.36 to -0.50), PROMIS physical health (r ranged from -0.39 to -0.58), and PROMIS mental health (r ranged from -0.37 to -0.57).				
Evans et al, 2012 (59)	✓✓ Pearson's correlation coefficient showed a significant association between the total scores on VOICE and the Service Satisfaction Scale: Residential measure (r = 0.82, p <0.001).	X	X	✓✓ Test–retest reliability was assessed with an interval of 6–10 days using Lin's concordance coefficient and was high (ρ = 0.88, CI = 0.81–0.95).	X
Forouzan et al, 2014 (60)	X	X	X	X	X
Garratt et al, 2006 (61) **Olsen et al, 2010 (62)**	X	X	X	✓ The test-retest reliability measured by the ICC was 0.90 for the POPEQ total scale and ranged from 0.75 to 0.89 for the subscales.	X
Gensichen et al, 2011 (63)	✓ Convergent validity was tested using the European	X	X	X	X

	Task Force on Patient Evaluations of General Practice (EUROPEP) questionnaire. Partial correlations ranged from 0.17 to 0.53 (all significant at $p<0.05$). Nevertheless, the two instruments have been designed by different methodological approaches, which may explain these moderate to low correlations.				
Gigantesco et al, 2003 (64)	X	X	X	✓✓ Test-retest reliability was evaluated by administering the questionnaire a second time within 4 to 6 days after the first time. The agreement between the two set was good, Cohen's weighted kappa ranging from 0.6 to 0.9 for most items and greater than 0.9 for some items. ICC was 0.80.	X
Glick et al, 1991 (65)	X	X	X	X	X
Hansson et al, 1995 (66)	X	X	X	X	X
Hester et al, 2015	✓ Convergent validity was established using the VSSS-	^X	X	✓ Test–retest reliability assessed over 5-7 days was	X

(67)	EU instrument, although it is a measure of user satisfaction rather than quality and is validated only for schizophrenic patients. Nevertheless, convergence was good (r=0.84, p < 0.001).			fair for the final 40-item version (r=0.65, p< 0.001).	
Howard et al, 2001 (68)	X	X	X	X	X
Ivarsson et al, 2007 (69) **Ahlfors et al, 2001 (70)**	✓ Correlations between Pat-UKU-ConSat and the original interviewer version ranged from 0.67 to 0.82 for single items, while correlations between sub-scores were respectively r=0.81 and r=0.75 (both significant at p<0.000001). Total score correlation was r=0.83.	X	X	X	X
Jenkinson et al, 2002 (71)	✓ Convergent validity was not assessed against a separate measure. The PPE-15 index was highly correlated with the total number of items selected as problems in the longer questionnaire (r ranged from 0.93 for Sweden to 0.95 for the other countries, p<0.001). These values exceeded the required cut-off of 0.9 for the development of short form	X	X	X	X

	measures.				
Joyce et al, 2010 (72)	X	X	✓✓ A series of hierarchical multiple linear regression analyses were conducted to assess the predictive validity. Perceived continuity was identified as significant predictor of generic health-related quality of life (EQ5D VAS, $p<0.002$), disease-specific quality of life (W-QLI weighted score, $p<0.0001$) and service satisfaction (SSS-30, $p<0.0001$). The observer-rated continuity appeared as a significant predictor only for the disease-related quality of life.	X	X
Kertesz et al, 2014 (73)	✓✓ Convergent validity was robust with a strong correlation between PCQ-H score and Roumie's single factor-derived score for the PCAS ($r=0.73$, $p<0.001$).	✓✓ As expected, the divergent validity was supported by a weak correlation with the Colorado symptom Index ($r=-0.13$, $p<0.002$)	X	X	X

Kolb et al, 2000 (74)	X	X	✓✓ The predictive validity of the instrument was evaluated using a stepwise regression model. All six scales showed significant criterion-related validity, explaining 58% of the variance in overall quality ratings (Pearson's correlation coefficient ranged from 0.43 to 0.67). Little shrinkage in this variance occurred when the model was cross-validated.	X	X
Larsen et al, 1979 (75)	✓ A correlation between therapists' estimates of the client's level of satisfaction with the actual client rating on the CSQ-8 provided some evidence of the scale's concurrent validity.	X	X	X	X
Lelliott et al, 2001 (76) **Blenkiron et al, 2003 (77)**	✓ CUES-U has been tested against the Health of the Nation Outcome Scales (HoNOS), although they are quite different in structure and mode of application. The total CUES-U score for part A is	X	X	✓ Test retest reliability was investigated by a second rating between 2 and 14 days after the first. ICCs were moderately good or better and ranged from 0.37-0.75 for part A and 0.52-	X

	significantly correlated with the total HoNOS score (r=0.42, p<0.01).			0.78 for part B.	
Llyod-Evans et al, 2010 (78)	X There is a lack of directly comparable instruments with demonstrated psychometric properties. Comparison with CaSPAR data did not establish convergent validity.	X	X	X	X
MacInnes et al, 2010 (79)	✓✓ Concurrent validity was supported by significant positive correlations (all at p<0.001) between the three subscales of the FSS and those corresponding to them in the VSSS (r ranged from 0.58 to 0.78) and for the total scores of the VSSS and the FSS (r=0.80).	X	X	X	X
Madan et al 2014 (80)	✓ Concurrent validity was established via significant correlations between MQOC total and two items that measure overall satisfaction: the assessment of overall quality (r=0.768, p<0.01) and likelihood of recommending (r=0.712, p<0.01).	X	X	X	X
Mavaddat et al, 2009 (81)	X	X	X	✓✓ Test-retest reliability estimates within 2–4 weeks	X

				after completing the first questionnaire was good (r=0.859, p=0.01).	
Mayston et al, 2017 (82)	X	X	✓ Predictive validity was examined using multiple logistic regressions adjusted for socio-demographic variables. Thus, the mean satisfaction score was significantly associated with four of the eight adherence items and two of the five therapeutic alliance items.	✓ Test-retest reliability estimates were assessed by the re-administration within 7–10 days of the first completion. Kappa's coefficient was lower than expected and ranged from -0.09 to 0.49, indicating agreement no better than that expected by chance.	X
McGuire et al, 2007 (83)	X	X	X	✓ Test–retest reliability evaluated over a 2-week interval was r=0.76 for the total STAR-P score and ranged from 0.68 to 0.81 for the three domains (all significant at p<0.05).	X
Meehan et al, 2002 (84)	X	X	X	X	X
Misdrahi et al, 2009 (85)	✓ Concurrent validity was assessed using a VAS added at the end of the 4PAS scale to estimate the level of alliance. A moderately high but	X	X	X	X

	significant correlation was obtained (r=0.62, p=0,00001).				
Moutoussis et al, 2000 (86)	X	X	X	X	X
Nabati et al, 1998 (87)	X	X	X	✓✓ Test-retest reliability assessed by the administration on two occasions, separated by 1-8 weeks, was supported by a significant correlation (r = 0.79, p = 0.05).	✓ The Satisfaction Index-Mental Health showed significant changes during clinical program change. Score increased significantly compared to baseline under treatment in the experimental easy-access program for bipolar disorder (47 + 12 vs. 59 + 11; Wilcoxon test, p<0.0001). Time in clinic dit not seem to have influenced this sensitivity to change (One-way ANOVA, p=0.90), as well as the regression to the mean.
Nordon et al, 2013 (88)	✓ Convergent validity was not assessed against another measure; however, some evidence of concurrent validity was provided and was fairly good. PASAP score was significantly associated to physician-rated satisfaction with life, illness severity, compliance, mood relapse and	X	X	X	✓ The responsiveness was assessed by computing effect sizes between 3 and 6 months with an effect size of at least 0.2 to support the sensitivity to change. The PASAP responsiveness was higher than the effect size of the other measured scales. However, the MCID of the PASAP score could not be

	adverse event.				tested using data of the EMBLEM study.
Oades et al, 2011 (42)	X	X	X	X	X
Rose et al, 2011 (89)	X	X	X	X	X
Ortiz et al, 2012 (90)	✓ Convergent validity was demonstrated by a significant positive correlation between each item and overall satisfaction measured from an anchor item (r ranging from 0.31 to 0.58). The total ICS scores and the sum of the scores of all items except the anchor item were also highly correlated (r=0.99).	X	✓✓ Predictive validity was examined using a multiple regression analysis. The correlation coefficients between each domain and the anchor item designated to measure overall care satisfaction ranged from 0.406 to 0.664 and was 0.707 for the total ICS scale.	X	X
Parker et al, 1996 (91)	X	X	X	X	X
Pellegrin et al, 2001 (92)	✓ Preliminary convergent validity was supported by significant correlations between all items and the scores of anchor that measured overall satisfaction with care (r=0.27 to 0.86 for the item "satisfaction with overall quality of care item" and r=0.19 to 0.63 for the item	X	X	X	X

	"likelihood of recommending the clinic to others item", p<0.05) and by significant correlations between the total scale scores and the scores for both anchor items (respectively, 0.80 and 0.63, p<0.05).				
Perreault et al, 2001 (93)	✓✓ Convergent validity was demonstrated by a strong correlation between OQOS-21 and CSQ-8 (r=0.77).	X	X	X	X
Perreault et al, 2006 (94)	✓✓ Convergent validity for the PPIQ-Satisfaction subscale with the CSQ-8 was established (r=0.50, p<0.01).	X	X	X	X
Peytremann-Bridevaux et al, 2006 (95)	X	X	X	X	X
Phattharayuttawat et al, 2005 (96)	X	X	X	✓✓ Test-retest reliability for total scale was high (0.82), as well as for each of the TPSS dimensions (ranged from 0.70-0.85).	X
Priebe et al, 1993 (97)	X	X	X	X	X
Rofail et al, 2005 (98)	X	X	X	X	X

Rose et al, 2009 (99)	X	X	X	✓ Test–retest reliability was measured by Lin's concordance coefficient with an interval of approximately 2 weeks between administrations. Lin's coefficient indicated substantial agreement for all dimensions and ranged from 0.73 to 0.82.	X
Røssberg et al, 2003 (100) **Røssberg et al, 2003 (101)**	X	X	X	X	X
Ruggeri et al, 2000 (102) **Ruggeri et al, 1993 (124)**	X	X	X	✓ Test-retest reliability in the pooled sample was high (0.82, $p<0.01$) but there were some differences between the sites (ICC ranged from 0.73 to 0.93, $p<0.01$). The ICC coefficients for the mean scores of each dimension ranged from 0.56 to 0.72 for the pooled sample, 0.49 to 0.72 for Amsterdam, 0.51 to 0.82 for Copenhagen, 0.66 to 0.82 for London, 0.43 to 0.80 for Santander and 0.51 to 0.76 for Verona.	X
Rush et al, 2013	✓✓ The correlation coefficient	X	X	X	X

(103)	between the total OPOC-MHA scores and CSQ-8 scale scores was moderate (r=0.58, p<0.001). Also, there is a good convergence between the four subscale scores and the CSQ-8 total scale score (Pearson's correlation coefficients ranged from 0.45 to 0.59, p<0.001). These moderate coefficients indicate that the OPOC-MHA covers broader perception of care aspects than the CSQ-8 does.				
Schalast et al, 2008 (104)	✓ Convergent validity was generally very good and was supported from significant correlations between EssenCES subscales and previously established scales. Thus, correlations were low to moderate for two subscales (r ranging from 0.25 to 0.78), while they were globally lower for the safety subscale (r ranging from 0.25 to 0.33).	✓There is a good evidence of the discriminant validity. As expected, it was supported by strong correlations between EssenCES scores by staff and patients' assessment of the ward climate (0.42–0.59) than with staff's job satisfaction (0.02–0.24).	X	X	X
Schröder et al, 2007 (105)	X	X	X	X	X
Schröder et al, 2010 (106)					
Schröder et al,					

2011 (107)					
Shiva et al, 2009 (108)	✓✓ Concurrent validity was established by moderate correlations between the revised-factor structure of the ISQ and the total and scale scores of the PCSQ and IESQ (all significant at p<0.01): the total ISQ score was slightly stronger correlated with the IESQ scale for both subsample (r=0.66 and r=0.67 respectively for forensic and civil sample), than with the total PCSQ score (r=0.60 and r=0.57 respectively).	X	X	X	X
Slade et al, 2014 (109)	✓ First, unlike what was expected, the unadjusted non-parametric correlations showed no association between Involvement subscale and any other variable. By contrast, both versions of Satisfaction subscale were associated with HAS-S (r ranged from 0.19 to 0.22, p<0.001) and HAS-P (r ranged from 0.36 to 0.42, p<0.001) and inversely associated with OQ-45 symptom distress (r ranged from -0.19 to -0.24, p<0.001). These relationships	✓✓ Divergent validity was established by no correlation between the two CDIS sub-scales and HoNOS, as well as no association between Involvement sub-scale and OQ-45 symptom distress.	✓✓ Predictive validity was investigated using ordinal regression models. For the CDIS-P, high satisfaction predicted decision implementation two months later for both continuous and utility versions of the Satisfaction sub-scale (respectively OR=2.21, p>0.001 and OR=3.13, p=0.033). In contrast, the Involvement subscale did not	X	X

	remained after adjusting for clustering by staff.		demonstrate predictive validity.		
Slater et al, 1982 (110)	X	X	X	✓ The SMHC scale was administered twice, with a 1 week interval between completions. All items had an ICC greater than 0.50 (data not shown).	X
Speckens et al, 2000 (111)	✓ There were strong significant correlations between the RQ scores and those on the Whitely Index (r ranged from 0.42 to 0.58) and the Health Anxiety subscale of the IAS (0.45 to 0.57) in all samples. Unlike was expected, the associations were weaker with the SSAS (0.25 to 0.41) and the Illness Behavior subscale scores of the IAS (0.24 to 0.42), and non-significant for the latter in the population 2 (r=0.11).	X The divergent validity was evaluated using the Kruskal-Wallis test to compare the scores in population 1, 2 and 3. The RQ scores were highest in the general population, as opposed to what was expected, but these differences were not significant between the three populations.	✓ High RQ scores tended to be predictive of a worse outcome in terms of recovery at 1-year follow-up ($p<0.05$). However, there was no relationship with the frequency of subsequent visits to the general practitioner (r=0.06).	✓✓ Test-retest reliability was assessed using Spearman rank correlation in population 3 and was 0.85 (ranged from 0.83 to 0.87).	✓ Sensitivity to change was assessed in a sample of 39 general medical outpatients with unexplained symptoms and treated with cognitive behavioral therapy by using the Wilcoxon matched-pairs signed-ranks test. RQ scores were lower at 6 months (Md=7, $p<0.05$) and 1-year follow-up (Md=4, $p<0.01$) than at baseline. The MCID was not computed.
Svedberg et al, 2008 (113) **Svedberg et al, 2007 (112)**	✓ Construct validity was established by significant positive correlations between HPIQ scores and those of the Empowerment scale (0.217, $p<0.05$), Client Satisfaction Scale (0.612, $p<0.001$) and the Helping Alliance Scale (0.762,	X	X	✓ Test-retest reliability was assessed using Cohen's kappa coefficient by administering the HPIQ twice with an interval of 4 weeks. It ranged from very good to poor (0.04 to 0.91). Thirteen out of 19 items had	X

	p<0.001). Contrary to what was expected, no associations were found between HPIQ scores and the Symptom Checklist.			moderate or higher reliability, while only one of them had poor reliability.	
Uijen et al, 2011 (114) **Uijen et al, 2012 (115)**	✓✓ Construct validity was demonstrated through expected correlations between NCQ score and scores of other variables. Thus, significant positive correlations (all significant at p<0.01) were found between the three subscales on general practice and the care suits patient subscale of the Consumer Quality Index (r=0.57–0.75), the general practice subscale of the VCC questionnaire (r=0.58–0.61), GP trust (r = 0.59–0.64) and satisfaction (r=0.63-0.67) scores. High correlations were also found between the three subscales on hospital/outpatient department care and the specialist care subscale of the VCC questionnaire (r=0.56–0.73), specialist trust (r=0.46–0.59) and satisfaction (r=0.48–0.59) scores. Finally, as expected, the team/cross-boundary continuity between GP and	X Discriminative ability was tested by investigating differences in continuity sub-scores of different subgroups.	X	✓✓ Test-retest reliability was calculated by having participants complete the NCQ a second time, 2 weeks after the initial completion. Reliability estimates was high (ICCs ranged from 0.71 to 0.82).	X Responsiveness has not been calculated, but the absence of a ceiling or floor effect is an indicator of the instrument's ability to show changes in scores over time.

	specialist subscale was at least moderately associated with the care suits patient subscale of the Consumer Quality Index (r=0.47), VCC subscales (r=0.56–0.65), trust subscales (r=0.27-0.30), and satisfaction subscales (r=0.33-0.38).				
Ul-Haq, 2012 (116)	X	X	X	X	X
Ware et al, 2003 (117)	✓Several evidence supporting convergent validity was provided in response to the lack of clear criteria for measuring continuity of care. To this end, associations with measures established in the literature representing the conceptual domains of clinical outcomes, quality of life, and satisfaction with care, were examined. The statistical associations were consistent with the hypothetical relationships: the strongest relationships were found with the indicators of the satisfaction domain. Also, there were low to modest statistically significant correlations between CONNECT scales and the clinical outcome and quality of	X	X	✓ Two-week test-retest reliability estimates indicated fair-to-good agreement for eleven scales (ICC ranged from 0.40 to 0.75). Two scales showed an excellent agreement (>0.75) and one scale showed a poor agreement (0.37).	X

	life indicators. As expected, no relationship was observed between the CONNECT scales and psychosis.				
Webster et al, 2012 (118)	X	X	X	X	X
Wongpakaran et al, 2013 (119)	✓✓ The GCS had a significant positive correlation with the Cohesion to Therapist Scale ($r=0.77$, $p<0.001$) and the Group Benefit Questionnaire ($r=0.71$, $p<0.001$).	X	✓ The level of prediction between items ranged from 27% to 69%.	X	X
Woodring et al, 2004 (120)	X	X	X	X	X
Wright et al, 2006 (125)	X	X	X	X	X
Zendjidjian et al, 2015 (121)	✓ The SATISPSY index was correlated with all the six VAS (r ranged from 0.41 to 0.62, $p<0.01$) and with each QSH dimensions (r ranged from 0.23 to 0.52, $p<0.01$). These low correlations with QSH seem to be due to specific concerns of psychiatric inpatients that the QSH generic satisfaction instrument does not cover.	✓ As expected, the SATISPSY-22 scores were globally more correlated with instrument scores of similar rather than dissimilar dimensions. Additionally, evidence of item-internal consistency and item discriminant validity was provided.	X	X	X

Zimmerman et al, 2017 (122)	✓ All items were significantly correlated ($p<0.001$) with each of the indicators of global satisfaction (an index of convergent validity). For the outpatient sample, correlations between all items and the overall level of satisfaction item ranged from 0.32 to 0.70 and those with the clinician's recommendation to friends/family item ranged from 0.22 to 0.70. For the partial hospital sample, correlations between each item and the overall level of satisfaction item ranged from 0.66 to 0.36.	✓ Higher correlations were found between items related to clinician behavior and overall satisfaction ratings than with those related to the office environment (a discriminant validity index).	X	X	X

✓✓ - positive rating, ✓ - acceptable rating, X - negative rating.