Chapter 28. Use of Structured Interviews, Rating Scales and Observational Methods in Clinical Settings

Argyris Stringaris, MD, PhD, FRCPsych
Faculty of Brain Sciences
University College London
&
First Dept. Psychiatry
National and Kapodistrian University of Athens

—First draft submitted by the author 26th April 2023—

Chapter 28. Use of Structured Interviews, Rating Scales and Observational Methods in Clinical Settings

Argyris Stringaris, University College London

28.0 Overview

The use of structured interviews, rating scales, and observational methods is now standard in child and adolescent psychiatry research and becoming more common in clinical practice. Here I provide a conceptual overview of these assessments, including their central assumptions and their challenges. I then discuss some major instruments used in the field. I will use the term *assessment instruments* to refer collectively to structured interviews, rating scales, and observational methods where necessary.

28.1 Background to Assessment Instruments

28.1.1 Importance of using assessment instruments

There are several reasons why using assessment instruments is important in clinical practice.

The first is evidence that using outcome measures can enhance the quality of care for young people. In this context, the term measurement-based outcomes or measurement-based care refers to employing standardised measures to regularly monitor a patient's symptoms, functioning, and treatment progress over time. Indeed, feeding back to clinicians on measurement appears to be useful. For example, Bickman and colleagues (Bickman et al., 2011) conducted a randomised cluster trial evaluated the effects of routine feedback to clinicians on mental health outcomes in youths receiving mental health services. They found that patients treated at sites where clinicians could receive weekly feedback improved faster than youths treated at sites where clinicians did not receive weekly feedback. Recent metanalyses further support the use of measurement feedback, though the effect sizes may be relatively modest (Cohen's *d* ~ 0.15)(Rognstad et al., 2023) and may work better when provided to both patients and clinicians (Edbrooke-Childs et al., 2015). Three are some similar findings specific to using standardised diagnostic instruments, there is evidence from a randomised controlled trial (RCT) that disclosing a standardised diagnosis to clinicians prior to clinical decision making improved agreement between clinicians for emotional disorders and led to the consideration of a broader range of diagnoses(Aebi et al., 2012); however, other trials found the effect to be modest and specific to some anxiety disorders(Ford et al., 2013). An ongoing trial examines in a large sample the impact that standardised diagnostic assessments have on the rate of diagnosis(Day et al., 2022).

The second is the imperative of integrating patients' views into the clinical decision making. It is important to think of the patient as an individual with unique needs, preferences, and values(Institute of Medicine (US) Committee on Quality of Health Care in America, 2001) and try to reflect this in the assessment process. Indeed, the importance of Patient Reported Outcome Measures (PROMS) has been repeatedly emphasised in the recent past and has found entry into government policy (NHS England, 2015) and is reflected in funders

expectations (Krause et al., 2021) in relation to child and adolescent mental health. Standardised measures are one (though not the only) way of incorporating patients' views into the clinical decision making. We elaborate on how the design of instruments and feedback is important in this regard.

The third concerns the utility of standardised instruments in freeing up clinician time. For example, asking patients to complete standardised measures before seeing them, can reduce the time that a clinician spends on screening for problems and thus free up time to focus on discussing with the patient what is most important to them. However, there is, to my knowledge, no empirical demonstration yet of this.

The fourth advantage concerns the ability to translate clinical constructs to quantities which can be statistically evaluated and compared with each other. This is a prerequisite for being able to offer patients evidence-based medicine.

All these are compelling reasons for using assessment instruments in clinical practice. Yet, there are several reasons to be cautious with their use and knowledgeable about their limitations. I expand on their conceptual, statistical and pragmatic challenges at the end of this chapter.

28.1.2 Assessment instruments as standardised measures of feelings and behaviour.

Structured interviews, rating scales, and observational methods are used increasingly in clinical practice. Their use is meant to allow clinicians to obtain "an objective and standardised measure of samples of behavior", as neatly summarised by two of the pioneers of psychometric assessment (Anastasi and Urbina, 1997) . These *samples of behaviour* can come in pretty much any format, including responses to a rating scale or observations made

by the clinician during parent-child interactions and they can include behaviours in the broad sense to include feelings and displays of emotion. The word *objective* can be confusing– here it should be taken to refer to the fact that the process and outcome of a test is empirically demonstrable and therefore subject to scrutiny (Urbina, 2016). They are *standardised* in two senses: first, that they are meant to be administered and scored in a uniform and systematic way to everybody (with exceptions in cases where linguistic, cultural or other adjustments are required). They are also standardised in the sense that their interpretation is based on some standard (e.g. performance of a population) and uses some criterion (e.g. total score on a scale)(Urbina, 2016).

The information derived by such instruments is then used for various clinical purposes, including in order to: arrive at a diagnosis, formulate a person's problems, develop a treatment plan, and monitor the child's progress over time.

28.1.3 Assumptions of assessment instruments and main methodological approaches

The assumption underlying assessment instruments is that there exists a true score, e.g. a true score of depression severity. The word *true* here does not need to imply that depression is a naturally-occuring entity, it is fine to consider this as a construct.

This assumption underlies both main approaches to assessment instruments, their design and interpretation, namely *Standard Test Theory (STT)* and *Item Response Theory (IRT)*.

Standard test theory (Lord et al., 1968) posits that a test score is composed of a true score and an error score. The true score is the hypothetical construct that represents an individual's actual level of ability or trait, e.g. depression severity. The error score reflects sources of fallibility such as factors associated with taking the test (e.g., being particularly tired on the day, or completing a questionnaire in a noisy environment) as well as problems with the test itself (e.g., poor wording or ambiguity of questions in depression). In Standard

Test Theory the aim is to minimise the impact of measurement error and obtain the most accurate estimate of the true score.

In contrast, IRT models (Embretson and Reise, 2000) the probability of an individual responding correctly to an item as a function of their ability level and the characteristics of the item, such as how difficult a question is to answer (e.g. the difficulty of a question is too vaguely or too precisely specified) and discrimination (how well an item differentiates between people with depression–questions about suicidality provide high discrimination). IRT assumes that an individual's level of ability is a continuous latent variable that is normally distributed across the population and models the measurement error as a function of the individual's level of ability. Overall, IRT provides a more nuanced approach to psychometric analysis than STT and has become increasingly popular in educational and psychological testing.

28.1.4 Basic psychometric parameters of assessment instruments

Assessment instruments are measures and this means that numbers are allocated to concepts (be it symptoms or syndromes) according to some rules that follow statistical principles. Here I discuss three key aspects of measures, namely reliability, validity and accuracy.

28.1.4.1 Reliability

When we measure an outcome, such as depression severity, in a group of people there is variability– outcomes vary between individuals. Reliability is central to psychometric theory and reflects the ratio of the true variability between people in the construct measured (e.g. depression severity) over the total variability of the measure (that includes the sources of

error mentioned above)(Nunnally and Bernstein, 1994). True variability cannot be measured directly and is therefore estimated in mainly two ways.

Cronbach's alpha: This is a measure of how strongly related the items of a measurement, e.g. different symptoms of depression, are with each other. When we measure an outcome, such as depression severity, in a group of people, each item on that scale will have a correlation with the other items. Cronbach's alpha is proportional to the average of those correlations as well as the total number of items in the scale and is inversely related to the total variability of the items. Values with a Cronbach's alpha of 0.8 and above are considered reliable, or internally consistent, though cut-offs should be applied with care.
There are two important things to know about Cronbach's alpha: first, a very high alpha can be obtained simply by creating very long scales (as intuited by the fact that alpha is proportional to the number of items). This is not only meaningless to strive for, but also a waste of precious patient time. Second, a high alpha can be obtained by asking the same question in various ways. This can create scales with high redundancy.

Test-retest reliability: This, most commonly assesses the consistency of scores obtained from the same individuals on the same instrument at two or more different points in time, as when the same patients are assessed on depression severity several days apart. It is easiest to intuit as the correlation between the same measure at the two different time points, although the use of different types of intra-class correlation coefficients is more appropriate depending on purpose. The higher the absolute test-reliability coefficient, the better, with values above 0.2 indicating at least fair reliability.
Figure 28.1 illustrates reliability using 10 clinic patients in four prototypical cases.

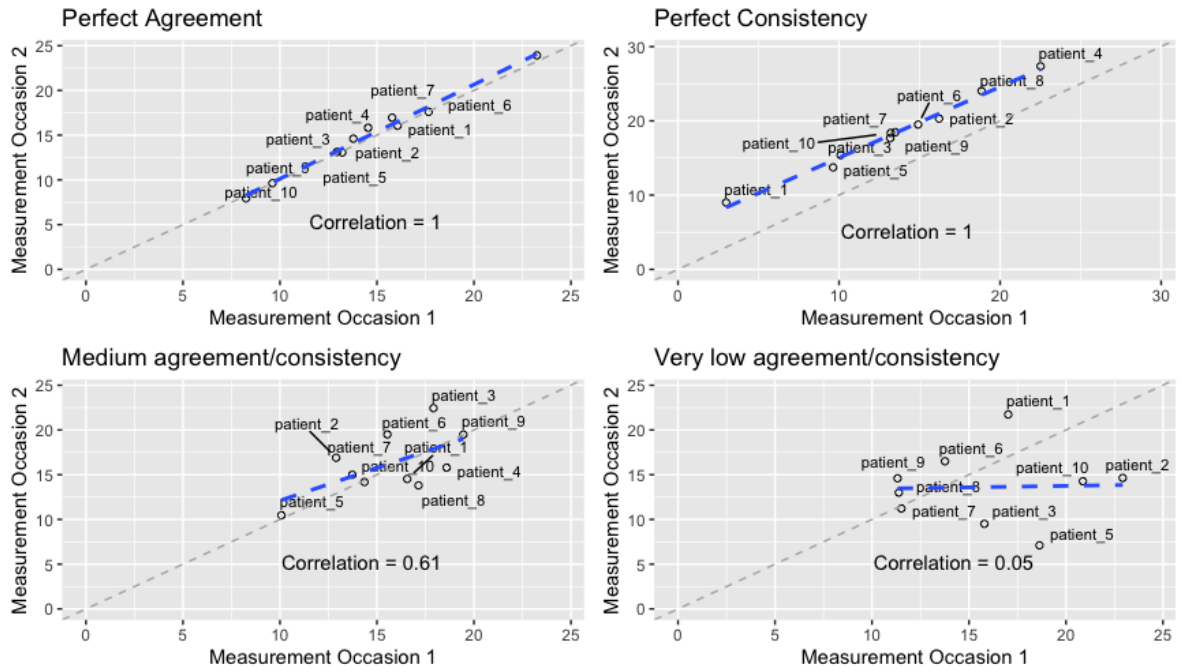


Figure 28.1 Illustrating patterns of reliability for two measurement occasions. On the top left-hand corner, an occasion of perfect agreement–in this fictitious (and highly unlikely) situation–patients score exactly the same on average on the measure on both occasions. Next to it, highly consistent measurements are displayed–every patient is shifted by 5 points between the two occasions (note that the correlation coefficient is the same as for the perfect agreement). On the bottom left, the more usual situation is depicted of medium agreement/consistency between the two measurement occasions. Next to it an instance of random ratings on the scale on each occasion. Data come from simulations by the author.

28.1.4.2 Validity

Validity in psychometrics refers to the extent to which an instrument measures what it is intended to measure. There are several types of validity that can be assessed to determine the extent to which an instrument is valid:

Content validity assesses the extent to which the items of a measure cover the aspects of interest to be measured. For example, does a depression questionnaire adequately represent the breadth of the experience of depression?

Face validity assesses whether the items of a measure are appropriate and relevant to the people who use them, for example, how well the items of a questionnaire ask about, say, low mood or anhedonia.

Face and content validity are increasingly recognised as key aspects of Patient Reported Outcome Measures (PROMs; see below) and increasing effort is being invested in taking them into account when creating instruments (Connell et al., 2018).

Criterion-related validity: This assesses how correlated a measure is with some criterion, or benchmark instrument of the construct being measured. In psychiatry, a standard interview or a previously used scale, are used as criteria against which a new scale will be measured. For example, one would expect any new measure of depression severity to demonstrate that it is strongly correlated with ones that are already in use.

Construct validity: This refers to the extent to which an instrument measures the construct it is intended to measure, rather than other constructs that may be related but distinct. For example, a questionnaire that purports to measure depression severity should correlate better with other scales that measure depression severity, than with scales that measure, say, ADHD.

28.1.4.3 Sensitivity, Specificity and Accuracy

In clinical practice, scales are often used to make predictions, such as assess membership into a group, e.g. does the score on this questionnaire indicate that the patient is suffering from depression or not. Below I discuss the parameters that are important in this regard based on a scenario.

Let there be a population where about half of young people have depression and the other half is healthy. You, the clinician, has two rating scales, A and B, that you are using in order

to screen this population with regards to depression. Obviously, your goal is to detect as many as possible of those who have depression and to avoid classifying healthy people as having depression. Figure 28.2 A shows the results received from using rating scale A in this population. Figure 28.2 B the results from rating scale B. It is obvious that the scores of rating scale A (the values on the y-axis of Figure 28.2 A, upper panel) is superior in discriminating between depressed and healthy adolescents: the questions it asks are more pertinent and that is why most people with depression have higher scores than healthy adolescents. By contrast, the scores on rating scale B (the values on the y-axis of Figure 28.2 B, upper panel) does not seem to distinguish between those depressed and those not. Rating scale A has better sensitivity and specificity, terms which I explain below. It also has better accuracy, as shown in the lower panel of each figure, another term that I explain below.

Sensitivity (also called true positive rate) refers to the ability of a test to identify correctly individuals who have the condition of interest, say how good a depression scale is at finding out if someone has depression. It is calculated as the proportion of people who are identified as having depression among all those who actually have depression (as measured against the gold standard of depression, e.g. an interview assessing depression). A high sensitivity is desired of screening instruments to make sure that cases are not missed.

Specificity (also called true negative rate), on the other hand, refers to the ability of a test or model to correctly identify individuals who do not have the condition of interest. In the case of depression, it would be the proportion of true negatives (i.e., individuals without depression who are correctly identified as such) amongst all people who do not have depression. A high specificity is desirable post-screening, particularly when making important treatment decisions.

Accuracy refers to the overall performance of a test or model, taking into account both sensitivity and specificity. In the depression case it would be the sum of true positives and true negatives divided over all the cases (true positives and negatives as well as false positives and negatives).

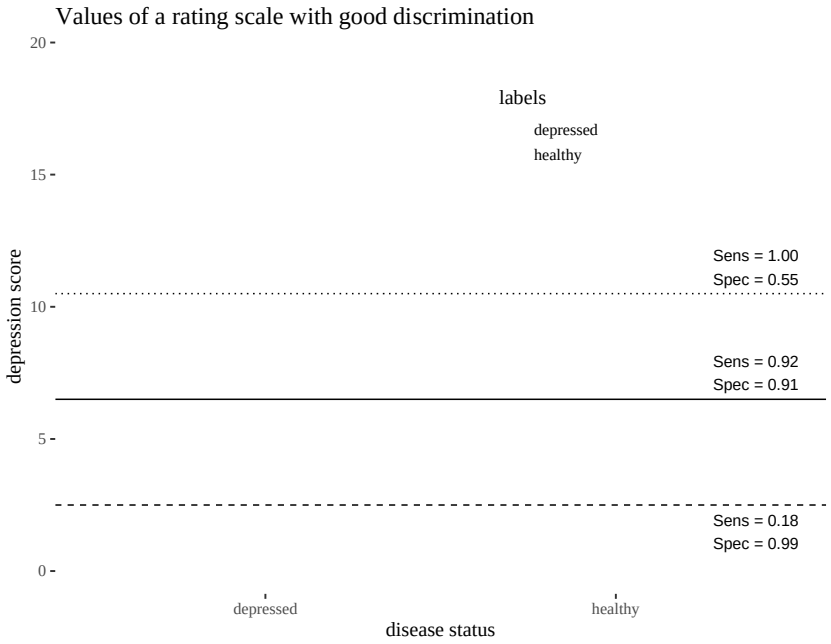
Area under the curve (AUC) refers to a graphical evaluation of the goodness of a binary (e.g. depression vs no depression). The idea is that one get a good sense of the relationship between how well a test detects (e.g. how well does it pick up depression) and also doesn't lead to too many false alarms (how often does it give us a false positive in depression). The lower panels of Figure 28.2, depict this relationship in a plot where the true positive rate (sensitivity) is plotted on the y-axis and the false positive rate (specificity reversely plotted) on the x-axis. This is, for historical reasons, called a Receiver Operator Curve (ROC). The test with the biggest area (AUC) under the ROC, is the best. An optimal test would be at (x,y) coordinates of 0,1 and it would have the maximal area; by contrast a test that performs as good as a coin flip, would be at 0.5,0.5, that is on the diagonal, the line that cuts the ROC in two triangular halves.

However, clinicians are also often than not faced with the reverse scenario, namely: if a test for depression is positive (or negative), what is the probability that their patient has (or, respectively, doesn't have) depression? The *positive predictive value* (PPV) is the ratio of True positives over all positive tests (true + false positives). This (and the equivalently calculated *negative predictive value*, NPV) are of practical relevance. The PPV and NPV are sensitive to the prevalence of the condition: the PPV is low when the prevalence is low: for example, a test for depression in a population where its prevalence is low, will yield high false positive rates and therefore a low PPV.

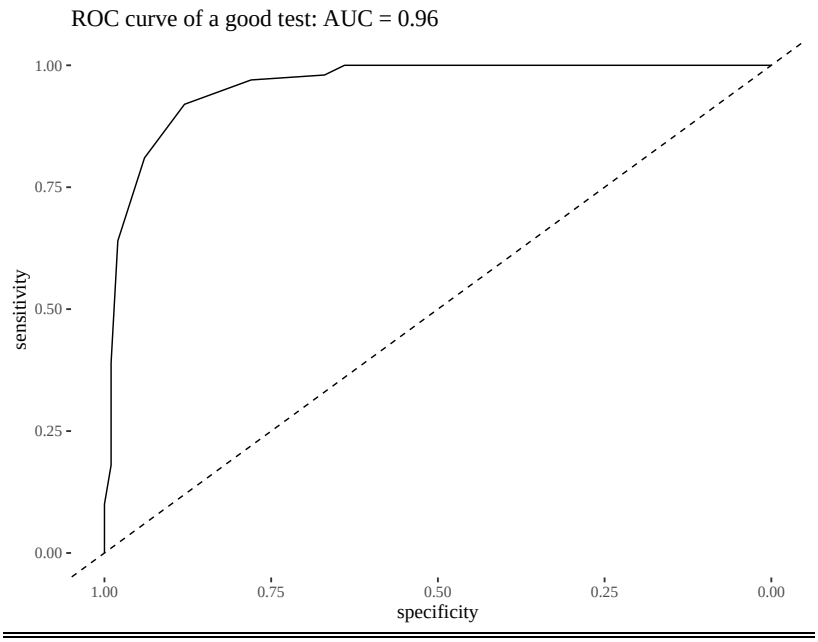
Figure 28.2 A tale of two rating scales.

A. Characteristics of a rating scale with good discrimination

28.2 Upper Panel

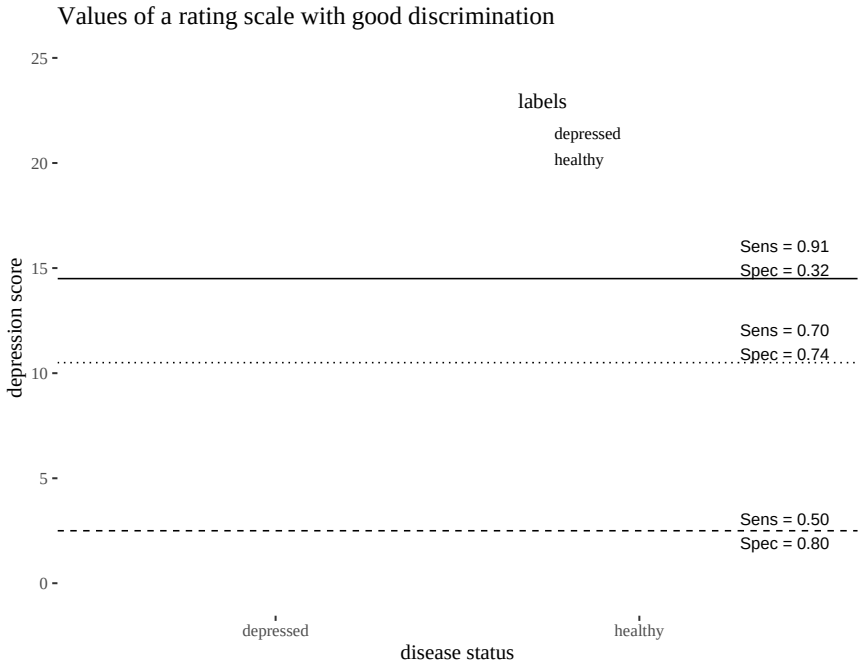


28.2 A Lower Panel



28.2 B. Characteristics of a rating scale with poor discrimination

28.2 B **Upper Panel**



28.2 B Lower Panel

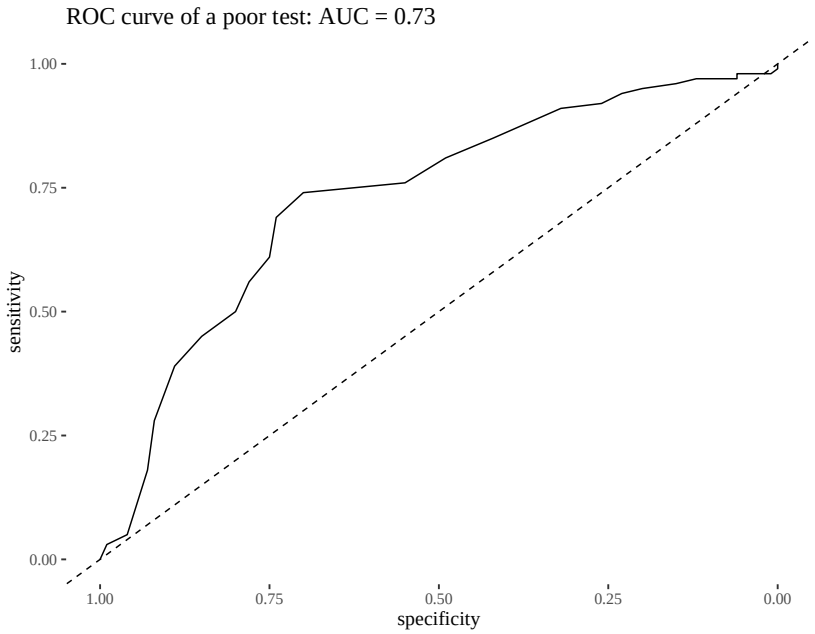


Figure 28.2 Here I illustrate sensitivity, specificity and AUC for a good (28.2 A) and a poor (28.2 B) rating scale.

Upper panel: The red dots are the values in a depression symptom questionnaire of patients with depression and the greed dots are the scores of healthy people. The black line is the value of the questionnaire (*threshold or cut-off score*) which maximises both sensitivity and specificity, as one moves upwards (e.g. dotted line) the sensitivity increases but the specificity decreases (many false positives, mislabelled healthy people), as one moves downwards (e.g. dashed line) the specificity increases at the cost of sensitivity (many false negatives, missed cases of depression).

Lower panel: AUC refers to the area under the curve (integral), for each rating scale showing rating scale A to be better than B. The dashed diagonal line is where one would expect chance detection. Figures are based on simulations run by the author (code liableble on request).

28.2 Types of Assessment Instruments

28.2.1 Interviews:

Semi-structured interviews are instruments that are typically conducted by clinicians (or researchers supervised by clinicians). They typically involve a set of standardised questions that the interviewer can use as a guide, but also allow for flexibility to explore areas of particular relevance to the individual being assessed. This type of interview is useful for gathering detailed information about specific symptoms or areas of concern, while still allowing for individual differences in presentation.

Structured interviews, on the other hand, are highly standardised and typically include a set of pre-determined questions and response options. This type of interview is useful for ensuring consistency in the assessment process across different patients and assessors, and can be particularly helpful for research purposes.

Most semi-structured and structured interviews follow either DSM-IV or DSM-5 criteria, some also ICD-10, to assess a wide range of mental health symptoms and disorders in children and adolescents, including anxiety, depression, ADHD, and autism spectrum disorder, among others. Please refer to **TABLE 1**, for some of the most widely used such instruments.

28.2.2 Rating scales

Rating scales provide a standardised way of measuring a range of mental health issues, such as anxiety, depression, attention-deficit/hyperactivity disorder (ADHD), and autism spectrum disorder (ASD), among others. Rating scales can be completed by various sources, including parents, teachers, the child, or adolescent themselves, and clinicians. Please refer to Table 1, for some of the most widely used rating scales for depression, anxiety, and behaviour problems. Rating scales typically enquire about symptoms or signs of problems–their intensity and/or frequency, as well as resulting impairments. Some scales explicitly enquire about strengths of the person. Some ratings scales also assess situational factors (e.g. when a problem arises) though this is the case less frequently.

There are broadband scales that enquire about psychopathology overall, with the SDQ and the CBCL being prime examples. These are used at assessment and sometimes at follow up.

There are specific scales used at assessment, such as the MFQ or SCARED for depression and anxiety, respectively.

Some scales are designed for assessing treatment progress, such as the CDRS-R, which has been modelled according to the adult HAM-D instrument.

28.2.3 Observational Measures

These measures involve the systematic observation and recording of a child's behavior, either through real-time observation or video recordings. Such measures can be used for diagnostic purposes, typically in ASD, or for the purposes of treatment planning and monitoring. Observations offer unique insights into the processes that may underlie the genesis or maintenance of psychopathology, such as parent-child conflict. They have been invaluable in understanding behavioural problems in children (such as oppositionality) and its treatment. They are particularly useful for very young children (e.g. mother-infant observations) and children who are limited in their verbal expression, and their use in core autism reflects such issues. Observational methods are limited in their utility by the effort and time they require to be set up, conducted and coded(Gardner, 2000). It is conceivable that modern technology, in particular automated software that can automatically decode affect and behaviour could allow for observational assessments at larger scales, though so far this promise remains unfulfilled (Dupré et al., 2020).

Perhaps the most widely used observational measure is the Autism Diagnostic Observation Schedule (ADOS). The ADOS contains both structured and semi-structured elements that assess the symptoms of autism spectrum disorder (ASD) in children and adolescents. It contains tasks and activities that assess social interaction, communication, and repetitive behaviours and is considered the gold standard for diagnosing ASD(Lord et al., 2000).

Other relatively widely used instruments include the Disruptive Behavior Diagnostic Observation Schedule (DB-DOS) a structured clinic-based assessment designed to elicit clinically salient behaviours relevant to the diagnosis of disruptive behaviour in pre-schoolers (Wakschlag et al., 2008).

Another observational measure is the Direct Observation Form (DOF), which is designed for rating observations of 6-11-year-old children in school classrooms, at recess, and in other group settings and can be used to assess classroom interventions (“Direct Observation Form (DOF)/Ages 6-11),” n.d.).

The Student Observation System (SOS) is another observation tool that is used to assess the behaviour of children and adolescents in the classroom setting. The SOS includes a set of standardised codes that capture behaviours such as academic engagement, disruptive behaviour, and social interactions(Kamphaus and Reynods, 2015).

Coding Interactive Behavior (CIB) is an observational measure used to assess the quality of parent-child interactions. The CIB includes a set of standardised codes that capture behaviours such as sensitivity, intrusiveness, and positive affect(Feldman, 2012).

28.2.4 Ecological momentary assessment (EMA)

EMA typically involves the use of electronic devices, such as smartphones or wearable sensors, to prompt participants to provide data multiple times a day, for several days or weeks. Young people are usually asked to respond to brief questionnaires or surveys, report their mood, activities, or events, or provide other types of information, such as physiological or GPS data.

The main advantage of EMA is that it provides a more detailed and accurate picture of people's behaviours and experiences in their everyday lives, compared to traditional methods that rely on retrospective self-reports (which are liable to several memory biases particularly over longer time frames) or laboratory settings. The promise of EMA is that it assesses the dynamic nature of mental health symptoms and behaviours in children and adolescents (Myin-Germeys et al., 2018). Here, dynamic refers to how symptoms change

over different time scales (e.g. how suicidal ideation fluctuates in real time) and how they may depend on environmental factors (e.g. how going to school or interacting with peers may affect depression severity) (Kleiman et al., 2017). Challenges to using EMA include: a) compliance and engagement as young, particularly those at school or in other activities may forget or be unable to respond to questions; b) technical issues, which include variations in young people's ability to use devices; c) ethical issues around data sharing but also potential over-exposure to electronic devices.

28.2.5 Chatbots

Chatbots are computer programmes that have been designed to simulate conversation with human users. Chatbots have been used for a variety of purposes, for both assessment and treatment delivery and may overall be acceptable to young people (Dosovitsky and Bunge, 2023), and also in trials(Nicol et al., 2022), though further research is required into their use (Abd-Alrazaq et al., 2020).

28.3 The basic ingredients of scale development and interpretation

In this section I discuss the importance of how standardised assessment instruments should take account of the perspectives of young people, describe briefly how scales developed (most recently in co-research with young people), and discuss two key components of their interpretation, namely the derivation of norms and standards, and the integration of different informant sources.

28.3.1 Lived experience, and youth goals and preferences.

Here I discuss the value of the perspective of people with lived experience in the design and application of assessment instruments and provide two examples of such instrument. Involving young people and families with lived experience in the design and development of assessment instruments is currently considered best practice for two main reasons (Scrutton, 2017). First, for reasons of epistemic justice, that is because it is unethical to ignore the perspective of the person with lived experiences themselves and devalue their knowledge about their own self. Second, for practical reasons, as it is expected to lead to more relevant, understandable, and acceptable assessment tools (i.e. improve both content and face validity, see above), that also empower young people in their mental health care. It will be interesting to also have empirical demonstration of how involving young people can lead to measurable improvements in healthcare.

It is relevant to discuss two types of instruments here that reflect the importance allocated to the views of health care users themselves. The first is *Preference based measurement*: this is in keeping with the idea of the importance of patients perspectives on their own health is a method used to measure the health-related quality of life (HRQoL) of children and adolescents. It is primarily used by economists to assesses individuals' preferences for different health states and allows for the calculation of quality-adjusted life years (QALYs), a measure of the value of a particular health state. A typical example of this is the Child Health Utility 9D (CHU9D) (Stevens and Ratcliffe, 2012). The second are *Goal-based measurement* which focuses on the achievement of specific goals that are collaboratively set between the patient and the provider. This approach is particularly important in child and adolescent psychiatry, where treatment goals may involve not only symptom reduction but also improvements in social, academic, and family functioning (Jacob et al., 2017). Goals are defined as working on the existing gap between current and desired states and the relevant changes in how one feels and acts that they want to achieve through therapy'(Jacob et al.,

2022, 2017). It involves questions such as: ‘what do you want to be different?’, ‘where do you want to go from here?’(Michalak and Holtforth, 2006).

Overall, people with lived experience can be involved as members of the research team, as advisors (e.g. youth advisory board) and as part of the team that disseminates outputs.

28.3.2 How are scales developed?

In order to understand the advantages and challenges of assessment instruments, it may help the clinical reader to have an overview of how scales are developed(Streiner et al., n.d.). Here I provide a simplified set of steps typically involved. I use the development of a depression measurement scale as an example.

Step 1: Conceptualization and Item Generation

The first step in developing a psychometric scale is to identify the construct that the scale will measure. In this case, the DSM-5 criteria for depression will be reviewed along with the literature of existing scales. Involving people with lived experience as researchers at this stage can be quite helpful in considering aspects of depression experience that academic researchers may not be aware of. It may also facilitate engagement with other young people with lived experience as research participants, and encourage them to offer views about the moods, feelings, cognitive and bodily experiences that are typically left out of scales. A recent historical review indicates how our current concept of depression may be unduly narrow(Kendler, 2016a). Based on this review, they will generate a pool of items that represent the various symptoms and features of depression, i.e. have face- and content-validity (see above).

Step 2: Item Selection

The next step is to select the most appropriate items from the pool generated in step one. This process involves a thorough review of each depression item, and items that are poorly

worded, redundant (e.g. asking about sadness in various different ways), or not relevant to the construct being measured are removed. The remaining items are then reviewed to ensure that they have good face validity and coverage of pertinent themes (e.g. anhedonia in depression). Again, involving young people at this stage can be very helpful and methodologies such as cognitive interviewing(Knafl et al., 2007) exist that provide a framework to scrutinise items for how understandable, acceptable and relevant they are.

Step 3: Scale Administration

Once the items have been selected, they are administered to a sample of individuals who have experience with the construct being measured, in this case, depression. The sample should be representative of the population of interest, and its size should be sufficient to provide adequate statistical power. Typically, respondents are asked to rate the extent to which they agree or disagree with each item using a Likert scale or another appropriate response format, such as a visual analogue scale.

Step 4: Analysis of acceptability and basic psychometric properties

The primary question is whether the scale is acceptable to people. One can ask young people directly about it, but rates of response can also be quite helpful in this. It is worth giving young people the opportunity to comment on individual items, as well as on the scale as a whole.

The data collected from the scale administration are analysed to evaluate the psychometric properties of the scale. This involves steps such as the following:

Factor Analysis: is used to identify the underlying structure of the scale items. In the case of a depression scale, whether the questions all measure one dimension (e.g. depression) or whether subscales can be identified (e.g. such that represent sadness vs loss of interest or somatic problems).

Reliability Analysis: this involves estimating Cronbach's alpha and test-retest reliability as described above.

Validity Analysis: Here the various forms of validity described above are estimated. Existing depression scales are used as outcomes (bearing in mind that very high coefficients of correlation may not be desired as this would indicate that a new scale might not be needed).

Step 5: Scale Revision and new data collection

Based on the results of the first-round data analysis, the scale may need to be revised. This involves removing or modifying items or delivery format and collect new data. Again, involving young people at each stage of the above process can help make scale development relevant to those who are meant to benefit from it.

Another important question about rating scales and more generally assessment instruments is whether they measure the same thing across time and across cultures. The frequency of behaviour/symptoms changes over time (e.g. think about temper outbursts) but their significance will also change over time (e.g. stealing at the age of 5, which may refer to taking out more chocolate than told to out of a bowl, vs stealing at the age of 16, which may refer to theft). Statistical approaches from the factor analysis tradition, chiefly *measurement invariance* evaluation, are primarily deployed to examine how much a test evaluates the same construct across time and place.

28.3.3 How are standards, norms and thresholds derived?

Using measured outcome to inform clinical decision making entails making normative judgements, that is, deciding what is bad or worse than before, and what is good or what may have gotten better. This is true for both outcome measured in a binary way (diagnosis or not) and those measured as a spectrum; this is because clinical decision making is categorical, e.g. treat or not, refer further or not, discharge or not, improved or not etc. Normative judgments are often implicit, they rely on societal expectations about age, gender and social position. They matter both in terms of how questions are asked and measures are designed, but also in how feedback is conveyed (Stringaris, 2021).

One way of developing norms is one based on what position a person occupies in a distribution of test scores. It involves collecting data from a representative sample of individuals who are similar to those for whom the test is intended, e.g. depressed adolescents. Once data are collected from the sample, the scores are analysed to determine the distribution of scores and the average score, in this case of depression. This then allows placing the individual on some position of the distribution of depression scores, e.g. finding out that their score corresponds to the 95th percentile of the population.

Another way of developing norms is one based on some criterion. For example, when validating an instrument for depression, one could use the diagnosis of depression as a criterion and establish the score that predicts depression with high enough accuracy. This is then used as a threshold for depression diagnosis and people can be above or below that criterion.

It should be obvious that in both examples above, the choice of the norm (the percentile or the diagnostic threshold) is an arbitrary one, it is not inherent in the quantitative measurement but rather a judgement that is influenced by values that prevail in society.

It is for this reason that the prevalence of disorders can vary substantially by where thresholds are set and what criteria are used for such thresholds(Costello et al., 2005).

Dealing with Informant Sources.

Child and adolescent mental health classically involves collecting information from various informants, including the young person, their parents, and teachers. The advantage is the richness of perspectives and the fact that disagreements about problems, their very existence or not, can be revealing. However, it also requires the ability by the clinician to integrate often disparate information. Research has consistently shown that there is often at best moderate agreement between informants about youth psychopathology. For example, a meta-analysis(Achenbach et al., 1987) found correlations between informants around r = 0.28 fairly modest agreement. Other studies have found similarly modest levels of agreement between informants on measures of depression, anxiety, and other mental health outcomes(De Los Reyes and Kazdin, 2005). There are several hypotheses about such

disagreements and they include measurement problems (e.g. the inability of young people to report reliably on their problems), limited knowledge (e.g. of parents and particularly teachers about a child's emotional problems, to which the child themselves have privileged access); the different setting at which children are observed, e.g. a child may behave differently at home than at school.

Another factor that may contribute to informant disagreement is the degree to which informants are aware of each other's reports. The disagreement between informants may also be simply an expression of the heterogeneity of a construct. For example, a recent study of informant sources of irritability showed substantial disagreement between parent and child reports of irritability despite good psychometric properties (Cronbach's alpha and test-retest reliability) for each informant report (Mallidi et al., 2023). This has led the authors to propose that children and parents interpret items differently and that this may be an indication of heterogeneity of the irritability construct.

In any case, clinicians should try to avoid simply averaging informant sources, as information may be lost. Similarly should also be cautious about simply applying an "or" rule to decide whether a diagnosis is present or not: again prevalence estimates can vary substantially depending on how informant reports are handled(Schwab-Stone et al., 1996). There is no recipe though on what to do with informant disagreement except perhaps to be open-minded about the possibility that either or both informants may be right in their rating (each from their own perspective perhaps). A clinician ought to document such differences and, where appropriate, try to resolve them in follow up interviews enquiring further with each informant or even by interviewing them together—differences in perspective are revealing for assessment and treatment

28.4 Practical Aspects of Assessment

28.4.1 Ethics and Conditions for Assessment

Ensuring confidentiality and to being upfront about what are the conditions of over-rule it (e.g. in the case of certain risks to self or others) is expected of every clinician conducting an assessment. Ensuring confidentiality can be tricky when conducting assessments remotely and when storing assessments online or electronically. Following best practices in the clinician's country and organisation is important and identifying information should be used sparingly and de-identification used whenever possible.
In general, the informant (patient or carer) should have easy access to the outcomes of the assessment and the assumption should be that the data belong to them.

The conditions should be conducive to assessment. This starts with "electronic convenience", the ease with which electronic assessments can be accessed and completed. There is some evidence that providing electronic platforms for completion of PROMS leads to higher completion rates compared to in-clinic completion(Morris et al., 2023). When in clinic, patients and carers should be given enough time and quiet space to complete their measures.
Clinicians should also be understanding if patients and carers have not had time to complete questionnaires–families with children with mental health difficulties will very often be disadvantaged in multiple other ways and short of time and resources (e.g. childcare). Making time to complete the questionnaires in clinic can be helpful.

28.4.2 How to choose assessment instruments

Different assessment instruments will be use for different purposes. In general, the best screening instruments will have a high sensitivity at the expense of specificity, wheres more specificity will be required when treatment decisions become necessary. Also, in the initial phases of assessment, broadband instruments (e.g. the Strengths and Difficulties Questionnaire(Goodman, 1997) will be more useful in getting a general sense of the breadth of problems, allowing the clinician to then hone in more at later stages using specific instruments, e.g. for depression or ADHD, at later stages.

Efficiency is important at every stage. Asking too many questions (having too long assessment batteries) risks attrition and dysphoria.

It is probably best for a service to have a standard battery of tests that everyone receives (e.g. a broadband scale) initially, which can then be refined, though this will also depend on the profile of the service–those that specialise in a certain condition, e.g. mood disorders will want to screen for that condition in sufficient depth in advance.

28.4.3 Considerations when collecting and interpreting data from assessment instruments

What is the issue?

The first question is what the problem is. As mentioned above, the leading problem may be different for children than for parents and carers. Simple open questions of the type: “I would like to hear from you about the reasons that we are meeting here/from your point of view, why are we meeting here today?” Instruments for assessing leading problems include the Top Problems tool, developed by Weisz and colleagues(Weisz et al., 2011), or Goal Based Outcomes(Jacob et al., 2022), particularly at the outset and monitoring of treatment.

Establishing a common language.

Parents and children often refer to the same problem in different ways. Neither patients nor their parents/ carers will necessarily use DSM-based terminology to communicate their problems. Words such as “bored” may be used to denote depressive feelings, whilst “I am naughty” may be the only thing that a child can say initially in order to describe their behaviour problems. It is important to pick up on such cues and enourage the person interviewed to tell you more about them during an interview. Both null findings as well as over-estimation of pathology in questionnaires can arise because the wording is outdated or inappropriate. Going through individual questions with the patient or carer can help, particularly when aspects of the assessment are known to be ambiguous.

Who reports the problem?

This will vary by the age of the children referred to the clinic, or the area of problems (e.g. OCD vs autism or ADHD) that the clinic specialises in. Both parents and teachers will be good at noticing the behaviours, and will therefore be good informants about behavioural equivalents of say irritability, i.e. the temper outbursts. However, particularly teachers may not be as good in describing accompanying feelings of anxiety or low mood, or the more persistent and often not externalised irritable mood. Typically-developing children as young as seven years, may provide meaningful accounts of their own feelings and of accompanying circumstances.

Placing the problem in a time frame.

In a semi-structured interview it is generally a good idea to elicit accounts of symptoms about specific events. Commonly this is achieved by asking the parents or young people to refer to “the most recent” or the “most severe” or the “one you can best remember”. If the problem is a daily occurrence, asking the person interviewed to take you through a “typical day” can be very helpful too. Using a whiteboard or sheet of paper to achieve this can help identify patterns. Trying to do this with both informant sources can be useful–disparate views on concrete events are revealing, as is the co-incidence of opinion.

Establishing a comparator.

When asking questions about the intensity, duration and consequences of actions, issues of normativity (see above and below) arise. This is typically dealt with by invoking a comparator, of which there are two principal ones.

The first comparator, used in many questionnaires and interviews asks the parent to rate children compared to other children of that age. This is particularly useful when asking about problems that occur repeatedly over time or are chronic.

The second comparator is the child themselves over time. This is the basis of assessing change over the last week or month. It is particularly useful when assessing episodic problems (e.g. depression) or those that have potentially sharply demarcated onsets (e.g. behavioural changes following abuse). The CDRS is an example of such an instrument in depression.

Intensity of the problem

When assessing intensity it is important to not simply rely on labels that the informant provides. Descriptions such as: “he is depressed” or “she had a tantrum” should be explored further and the clinician should look for individual signs or symptoms through their interview. Questionnaires serve the same purpose by asking about several possible symptoms.

The clinician will look at the report of the parent, carer or teacher for signs of distress. For example, the following information is useful in judging a child’s tantrum: the extent of angry facial expressions, being red in the face, yelling, screaming, and stamping their feet. Breaking things, attacking others is an obvious sign of escalation and higher intensity. Similarly, facial expressions of sadness, the intensity of crying, and the degree to which a person is unresponsive to pleasant stimuli are all information relevant to the assessment of depression.

Looking at the total score of questionnaires is important but can also be misleading. Certain conditions, such as Obsessive Compulsive Disorder or Body Dysmorphic Disorder may be very impairing even when monosymptomatic–one single symptom such as hand washing can be so frequent or of such long duration as to be devastating.

Chronicity

The time of onset of a problem is not always clear and research instruments may avoid such questions to avoid recall biases. However, hearing about onset can be quite helpful for understanding perceptions of the problem. Characteristically in the case of irritability, one may receive answers from the parent of the type: “forever, ever since he was born”. Similarly, in depression the time of onset may be blurry. By contrast, there may be very clear and sharply demarcated onsets, and these can give clues about important life events, such as trauma.

Clinicians will want to know whether depression, anxiety or irritability occurs as discrete episodes—e.g. a period of one week of very intense low mood or irritability—or as something chronic, that is, a set of feelings and behaviours that have been typical of the young person for a considerable period of time. Drawing a timeline—perhaps on a board or a large sheet of paper —can be quite helpful.

Context & antecedents

Some symptoms will manifest in certain circumstances only, such as when reminded of traumatic events or when confronted with particular challenges, such as when a child with autism is faced with un-announced change. Other symptoms are pervasive, i.e. they can occur at home, as much as at school or during leisure activities. Pervasiveness is generally regarded as an indicator of severity, however, symptoms that occur only in particular contexts can also be debilitating, characteristically oppositionality that is restricted to the parents. Conventional questionnaires and interviews are often not designed to assess this and thus deprive the clinician of important information. Diary keeping and recording by tablets or smartphones can be particularly helpful, provided that informants are sufficiently familiar with them. This is a domain where mobile technologies and EMA can be particularly useful.

Impact and Consequences.

It is worth thinking about these problems along two axes.

The first axis concerns the domains or environment in which problems manifest. Often these are divided into family life, school environment (to include academic performance and behaviour), leisure time and peer relationships. This division of domains is reflected in several instruments, including in the impact supplement of the SDQ.
The second axis concerns the time scale of impact and consequences and is typically divided into immediate (or short term), medium- and long-term consequences.
There are several scales one can use for assessing functioning and impact, including the clinician-rated Children's Global Assessment Scale (CGAS)(Shaffer et al., 1983) and the parent-, teacher- or youth reported SDQ impact supplement.

28.5 Limitations and criticisms concerning Assessment Instruments.

As mentioned at the beginning of this chapter, the use of assessment instruments offers several advantages and is becoming more common in clinical practice. However, there are several reasons to be cautious in the interpretation of results from assessment instruments. I enumerate these below.

Undue Reification: It is common to view psychiatric diagnoses or dimensions as real entities, as if our classification had carved nature at its joints(Hoff, 2017). This is very unlikely to be the case for most psychiatric and psychological disorders, and alternative explanations have been proposed(Kendler, 2016b). In any case, it is important to display sufficient epistemic humility when referring to diagnoses and dimensions arrived at through assessment instruments: it is very unlike that, at least at the moment, our measurement in psychiatry resembles that of, say, temperature measurement using mercury.

Narrow representation: It is striking that most assessment instruments have been developed in a subset of Western Industrialised countries. Many of these instruments have been translated and applied to populations outside those countries, but the fact that they are inspired and driven primarily by concerns in one part of the world, may constrain our knowledge about variation in psychological phenomena. It may lead to either over- or under-recognition of problems in people. Perhaps inevitably, most instruments are developed within the bounds of the normativities (Stringaris, 2021) of a certain historical epoch and of a dominant class of people within a society. Such normativities shape how instruments are developed and what is measured. As a result, experiences of those less represented, including those of different sexual orientation or gender minorities, may be absent. This is a particular problem given the disproportionately higher rates of psychopathology in certain under-represented groups (Reisner et al., 2016). Similar issues arise when applying criteria developed for boys (as in ADHD research(Hinshaw et al., 2022)) onto girls, or vice versa and when symptom threshold are applied uncritically across age ranges.

Psychometric criticism: Criticisms that apply to the DSM or ICD classification systems also apply to assessment instruments. For example, based on current criteria, one can arrive at thousands of different depression syndromes, simply based on combinatorics, and this heterogeneity is also demonstrable empirically(Fried and Nesse, 2015). Similarly, instruments are developed in ways that are not always transparent, some items are simply used ad hoc, decisions about the inclusion or exclusion of items and the statistics are often not clear, and several instruments may be developed that purport to measure the same thing, but measure something different, or purport to measure different things, but actually show great overlap(Flake and Fried, 2020).

Access: Many instruments are behind paywalls, even though public funds have been used (sometimes exclusively) for their development and despite the fact that the income from such instruments is reasonably thought to far exceed maintenance and related costs. This is

particularly frustrating when it concerns instruments that have become “gold standard” as it means that cash-strapped clinics and health care systems will not have access to them and that patients will not benefit from their use.

Table 28.1 Common interviews for general psychopathology in child and adolescent psychiatry.

Kiddie Schedule for Affective Disorders and Schizophrenia (K-SADS)	Semi-structured diagnostic interview used to assess current and past episodes of psychopathology in children and adolescents.	(Kaufman et al., 1997)
Diagnostic Interview Schedule for Children (DISC)	Structured diagnostic interview used to assess a wide range of mental disorders in children and adolescents.	(Shaffer et al., 1996)
Development and Well-Being Assessment (DAWBA)	Structured diagnostic interview used to assess a wide range of mental	(Goodman et al., 2000)

	disorders in children and adolescents.	
Child and Adolescent Psychiatric Assessment (CAPA)	Structured diagnostic interview used to assess a wide range of psychiatric disorders and substance abuse in children and adolescents.	(Angold et al., 1995)

Table 28.2. Common rating scales in child and adolescent psychiatry

Assessment Instrument	Description	Reporter(s)	No. of Items	Reference
Strengths and Difficulties Questionnaire	Broadband psychopathology	self, parent, teacher	25	(Goodman, 1997)
Child Behavior Checklist (CBCL)	Broadband psychopathology	Parent or self-report	113	(Achenbach and Rescorla, 2004)
Beck Depression Inventory (BDI-II)	Measures severity of depressive symptoms	Self	21	(Beck et al., 1996)

Children's Depression Inventory (CDI)	Measures depressive symptoms in children and adolescents	Self, Parent	27	(Kovacs, 1985)
Children's Depression Rating Scale-Revised	Measures severity of depressive symptoms in children and adolescents	Clinician, Parent	17	(Poznanski et al., 1984)
Center for Epidemiological Studies Depression Scale (CES-D)	Measures depressive symptoms in the general population	Self	20	(Garrison et al., 1991)
Reynolds Adolescent Depression Scale (RADS-2)	Measures depressive symptoms in adolescents	Self, Clinician	30	(Reynolds, 2004)
Mood and Feelings Questionnaire (MFQ)	Measures depressive symptoms in children and	Self, Parent	33	(Angold et al., 1995)

	adolescents			
Revised Children's Anxiety and Depression Scale (RCADS)	Measures anxiety and depressive symptoms in children and adolescents	Self, Parent	47	(Chorpita et al., 2000)
Screen for Child Anxiety Related Emotional Disorders (SCARED)	Measures anxiety symptoms in children and adolescents	Self, Parent	41	(Birmaher et al., 1997)
Spence Children's Anxiety Scale (SCAS)	Measures anxiety symptoms in children and adolescents	Self, Parent	44	(Spence et al., 2003)
Multidimensional Anxiety Scale for Children (MASC)	Measures anxiety symptoms in children and adolescents	Self, Parent	39	(March et al., 1997)

Conners' Rating Scales-Revised (CRS-R)	Measures symptoms of ADHD	Parent, Teacher	27	(Conners et al., 1998)
SNAP-IV ADHD	Measures symptoms of ADHD	Parent, Teacher	26	(Swanson et al., 2001)
Affective Reactivity Index (ARI)	Irritability: preschool to age 18 years	Parent or Self	7 items	(Stringaris et al., 2012)
Clinician Affective Reactivity Index (CL-ARI)	Irritability: 6 to age 18 years	Clinician	12	(Haller et al., 2020)
Emotion Dysregulation Inventory	Emotion dysregulation with subscales for reactivity and dysphoria	Parent	30	(Mazefsky et al., 2018)

Social Communications Questionnaire	Assessing social communication skills in ASD	Parent	40	(Rutter et al., 2003)
Social Responsiveness Scale (SRS)	Assessing social responsiveness in ASD	Parent or teacher	65	(Constantino and Gruber, 2005)

REFERENCES

Angold, A., Prendergast, M., Cox, A., Harrington, R., Simonoff, E., Rutter, M., 1995. The Child and Adolescent Psychiatric Assessment (CAPA). Psychological Medicine 25, 739–753. https://doi.org/10.1017/S003329170003498X

Constantino, J., Gruber, C., 2005. The Social Responsiveness Scale. Western Psychological Sciences, Los Angeles.

Hinshaw, S.P., Nguyen, P.T., O'Grady, S.M., Rosenthal, E.A., 2022. Annual Research Review: Attention-deficit/hyperactivity disorder in girls and women: underrepresentation, longitudinal processes, and key directions. J Child Psychol Psychiatry 63, 484–496. https://doi.org/10.1111/jcpp.13480

Kamphaus, R., Reynods, C., 2015. BASC-3 | The Behavior Assessment System for Children [WWW Document]. URL https://www.pearsonassessments.com/store/usassessments/en/Store/Professional-Assessments/Behavior/BASC-3-Family-of-Assessments/p/P100010000.html?gclid=CjwKCAjw586hBhBrEiwAQYEnHYy-voOJRDXZFG5YPxGKvvRIpEqievQmm9OzcKmV3mhmqzCOmKt8bxoCmMQQAvD_BwE (accessed 4.10.23).

NHS England, 2015. Future in Mind: Children and Young People's Mental Wellbeing [WWW Document]. URL https://www.england.nhs.uk/blog/martin-mcshane-14/ (accessed 4.9.23).

Rutter, M., Bailey, A., Lord, C., 2003. The Social Communication Questionnaire. Western Psychological Sciences, Los Angeles.

Abd-Alrazaq, A.A., Rababeh, A., Alajlani, M., Bewick, B.M., Househ, M., 2020. Effectiveness and Safety of Using Chatbots to Improve Mental Health: Systematic Review and Meta-Analysis. Journal of Medical Internet Research 22, e16021. https://doi.org/10.2196/16021

Achenbach, T.M., McConaughy, S.H., Howell, C.T., 1987. Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. Psychological Bulletin 101, 213–232. https://doi.org/10.1037/0033-2909.101.2.213

Achenbach, T.M., Rescorla, L.A., 2004. The Achenbach System of Empirically Based Assessment (ASEBA) for Ages 1.5 to 18 Years, in: The Use of Psychological Testing for Treatment Planning and Outcomes Assessment: Instruments for Children and Adolescents, Volume 2, 3rd Ed. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US, pp. 179–213.

Aebi, M., Kuhn, C., Metzke, C.W., Stringaris, A., Goodman, R., Steinhausen, H.-C., 2012. The use of the development and well-being assessment (DAWBA) in clinical practice: a randomized trial. Eur Child Adolesc Psychiatry 21, 559–567. https://doi.org/10.1007/s00787-012-0293-6

Anastasi, A., Urbina, S., 1997. Psychological testing, 7th ed, Psychological testing, 7th ed. Prentice Hall/Pearson Education, Upper Saddle River, NJ, US.

Angold, A., Costello, E.J., Messer, S.C., Pickles, A., 1995. Development of a short questionnaire for use in epidemiological studies of depression in children and adolescents. International Journal of Methods in Psychiatric Research 5, 237–249.

BASC-3 | The Behavior Assessment System for Children, Third Edition [WWW Document], n.d. URL https://www.pearsonassessments.com/store/usassessments/en/Store/Professional-Assessments/Behavior/BASC-3-Family-of-Assessments/p/P100010000.html?gclid=CjwKCAjw586hBhBrEiwAQYEnHYy-voOJRDXZFG5YPxGKvvRIpEqievQmm9OzcKmV3mhmqzCOmKt8bxoCmMQQAvD_BwE (accessed 4.10.23).

Beck, A.T., Steer, R.A., Ball, R., Ranieri, W., 1996. Comparison of Beck Depression Inventories -IA and -II in psychiatric outpatients. J Pers Assess 67, 588–597. https://doi.org/10.1207/s15327752jpa6703_13

Bickman, L., Kelley, S.D., Breda, C., de Andrade, A.R., Riemer, M., 2011. Effects of routine feedback to clinicians on mental health outcomes of youths: results of a randomized trial. Psychiatr Serv 62, 1423–1429. https://doi.org/10.1176/appi.ps.002052011

Birmaher, B., Khetarpal, S., Brent, D., Cully, M., Balach, L., Kaufman, J., Neer, S.M., 1997. The Screen for Child Anxiety Related Emotional Disorders (SCARED): scale construction and psychometric characteristics. J Am Acad Child Adolesc Psychiatry 36, 545–553. https://doi.org/10.1097/00004583-199704000-00018

Chorpita, B.F., Yim, L., Moffitt, C., Umemoto, L.A., Francis, S.E., 2000. Assessment of symptoms of DSM-IV anxiety and depression in children: A revised child anxiety and depression scale. Behav Res Ther 38, 835–855. https://doi.org/10.1016/s0005-7967(99)00130-8

Connell, J., Carlton, J., Grundy, A., Taylor Buck, E., Keetharuth, A.D., Ricketts, T., Barkham, M., Robotham, D., Rose, D., Brazier, J., 2018. The importance of content and face validity in instrument development: lessons learnt from service users when developing the Recovering Quality of Life measure (ReQoL). Qual Life Res 27, 1893–1902. https://doi.org/10.1007/s11136-018-1847-y

Conners, C., Sitarenios, G., Parker, J., Epstein, J., 1998. Conners CK, Sitarenios G, Parker JD, Epstein JN. The revised Conners' Parent Rating Scale (CPRS-R): factor structure, reliability, and criterion validity. J Abnorm Child Psychol 26: 257-268. Journal of abnormal child psychology 26, 257–68. https://doi.org/10.1023/A:1022602400621

Costello, E.J., Egger, H., Angold, A., 2005. 10-Year Research Update Review: The Epidemiology of Child and Adolescent Psychiatric Disorders: I. Methods and Public Health Burden. Journal of the American Academy of Child & Adolescent Psychiatry 44, 972–986. https://doi.org/10.1097/01.chi.0000172552.41596.6f

Day, F., Wyatt, L., Bhardwaj, A., Dubicka, B., Ewart, C., Gledhill, J., James, M., Lang, A., Marshall, T., Montgomery, A., Reynolds, S., Sprange, K., Thomson, L., Bradley, E., Lathe, J., Newman, K., Partlett, C., Starr, K., Sayal, K., 2022. STAndardised DIagnostic Assessment for children and young people with emotional difficulties (STADIA): protocol for a multicentre randomised controlled trial. BMJ Open 12, e053043. https://doi.org/10.1136/bmjopen-2021-053043

De Los Reyes, A., Kazdin, A.E., 2005. Informant Discrepancies in the Assessment of Childhood Psychopathology: A Critical Review, Theoretical Framework, and Recommendations for Further Study. Psychological Bulletin 131, 483–509. https://doi.org/10.1037/0033-2909.131.4.483

Direct Observation Form (DOF)/Ages 6-11) [WWW Document], n.d. . ASEBA. URL https://aseba.org/direct-observation-form-dof-ages-6-11/ (accessed 4.10.23).

Dosovitsky, G., Bunge, E., 2023. Development of a chatbot for depression: adolescent perceptions and recommendations. Child and Adolescent Mental Health 28, 124–127. https://doi.org/10.1111/camh.12627

Dupré, D., Krumhuber, E.G., Küster, D., McKeown, G.J., 2020. A performance comparison of eight commercially available automatic classifiers for facial affect recognition. PLOS ONE 15, e0231968. https://doi.org/10.1371/journal.pone.0231968

Edbrooke-Childs, J., Jacob, J., Law, D., Deighton, J., Wolpert, M., 2015. Interpreting standardized and idiographic outcome measures in CAMHS: What does change mean and how does it relate to functioning and experience? Child Adolesc Ment Health 20, 142–148. https://doi.org/10.1111/camh.12107

Embretson, S.E., Reise, S.P., 2000. Item Response Theory. Psychology Press, New York. https://doi.org/10.4324/9781410605269

Feldman, R., 2012. Parenting behavior as the environment where children grow, in: The Cambridge Handbook of Environment in Human Development, Cambridge Handbooks in Psychology. Cambridge University Press, New York, NY, US, pp. 535–567. https://doi.org/10.1017/CBO9781139016827.031

Flake, J.K., Fried, E.I., 2020. Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them. Advances in Methods and Practices in Psychological Science 3, 456–465. https://doi.org/10.1177/2515245920952393

Ford, T., Last, A., Henley, W., Norman, S., Guglani, S., Kelesidi, K., Martin, A.-M., Moran, P., Latham-Cork, H., Goodman, R., 2013. Can standardized diagnostic assessment be a useful adjunct to clinical assessment in child mental health services? A randomized controlled trial of disclosure of the Development and Well-Being Assessment to practitioners. Soc Psychiatry Psychiatr Epidemiol 48, 583–593. https://doi.org/10.1007/s00127-012-0564-z

Fried, E.I., Nesse, R.M., 2015. Depression is not a consistent syndrome: An investigation of unique symptom patterns in the STAR*D study. J Affect Disord 172, 96–102. https://doi.org/10.1016/j.jad.2014.10.010

Gardner, F., 2000. Methodological Issues in the Direct Observation of Parent–Child Interaction: Do Observational Findings Reflect the Natural Behavior of Participants? Clin Child Fam Psychol Rev 3, 185–198. https://doi.org/10.1023/A:1009503409699

Garrison, C.Z., Addy, C.L., Jackson, K.L., McKEOWN, R.E., Waller, J.L., 1991. The CES-D as a Screen for Depression and Other Psychiatric Disorders in Adolescents. Journal of the American Academy of Child & Adolescent Psychiatry 30, 636–641. https://doi.org/10.1097/00004583-199107000-00017

Goodman, R., 1997. The Strengths and Difficulties Questionnaire: A Research Note. Journal of Child Psychology and Psychiatry 38, 581–586. https://doi.org/10.1111/j.1469-7610.1997.tb01545.x

Goodman, R., Ford, T., Richards, H., Gatward, R., Meltzer, H., 2000. The Development and

Well-Being Assessment: description and initial validation of an integrated assessment of child and adolescent psychopathology. J Child Psychol Psychiatry 41, 645–655.

Haller, S.P., Kircanski, K., Stringaris, A., Clayton, M., Bui, H., Agorsor, C., Cardenas, S.I., Towbin, K.E., Pine, D.S., Leibenluft, E., Brotman, M.A., 2020. The Clinician Affective Reactivity Index: Validity and Reliability of a Clinician-Rated Assessment of Irritability. Behav Ther 51, 283–293. https://doi.org/10.1016/j.beth.2019.10.005

Hoff, P., 2017. On reification of mental illness: Historical and conceptual issues from Emil Kraepelin and Eugen Bleuler to DSM-5, in: Philosophical Issues in Psychiatry IV: Classification of Psychiatric Illness, editors: Kenneth Kendler, Josef Parnas. Oxford University Press.

Institute of Medicine (US) Committee on Quality of Health Care in America, 2001. Crossing the Quality Chasm: A New Health System for the 21st Century. National Academies Press (US), Washington (DC).

Jacob, J., Edbrooke-Childs, J., Flannery, H., Segal, T.Y., Law, D., 2022. Goal-based measurement in paediatric settings: implications for practice. Archives of Disease in Childhood. https://doi.org/10.1136/archdischild-2021-322761

Kaufman, J., Birmaher, B., Brent, D., Rao, U., Flynn, C., Moreci, P., Williamson, D., Ryan, N., 1997. Schedule for Affective Disorders and Schizophrenia for School-Age Children-Present and Lifetime Version (K-SADS-PL): initial reliability and validity data. J Am Acad Child Adolesc Psychiatry 36, 980–988. https://doi.org/10.1097/00004583-199707000-00021

Kendler, K.S., 2016a. The Phenomenology of Major Depression and the Representativeness and Nature of DSM Criteria. Am J Psychiatry 173, 771–780. https://doi.org/10.1176/appi.ajp.2016.15121509

Kendler, K.S., 2016b. The nature of psychiatric disorders. World Psychiatry 15, 5–12. https://doi.org/10.1002/wps.20292

Kleiman, E.M., Turner, B.J., Fedor, S., Beale, E.E., Huffman, J.C., Nock, M.K., 2017. Examination of real-time fluctuations in suicidal ideation and its risk factors: Results from two ecological momentary assessment studies. J Abnorm Psychol 126, 726–738. https://doi.org/10.1037/abn0000273

Knafl, K., Deatrick, J., Gallo, A., Holcombe, G., Bakitas, M., Dixon, J., Grey, M., 2007. The analysis and interpretation of cognitive interviews for instrument development. Res Nurs Health 30, 224–234. https://doi.org/10.1002/nur.20195

Kovacs, M., 1985. The Children's Depression, Inventory (CDI). Psychopharmacol Bull 21, 995–998.

Krause, K.R., Chung, S., Adewuya, A.O., Albano, A.M., Babins-Wagner, R., Birkinshaw, L., Brann, P., Creswell, C., Delaney, K., Falissard, B., Forrest, C.B., Hudson, J.L., Ishikawa, S.-I., Khatwani, M., Kieling, C., Krause, J., Malik, K., Martínez, V., Mughal, F., Ollendick, T.H., Ong, S.H., Patton, G.C., Ravens-Sieberer, U., Szatmari, P., Thomas, E., Walters, L., Young, B., Zhao, Y., Wolpert, M., 2021. International consensus on a standard set of outcome measures for child and youth anxiety, depression, obsessive-compulsive disorder, and post-traumatic stress disorder. Lancet Psychiatry 8, 76–86. https://doi.org/10.1016/S2215-0366(20)30356-4

Lord, C., Risi, S., Lambrecht, L., Cook, E.H., Leventhal, B.L., DiLavore, P.C., Pickles, A., Rutter, M., 2000. The Autism Diagnostic Observation Schedule—Generic: A Standard Measure of Social and Communication Deficits Associated with the Spectrum of Autism. J Autism Dev Disord 30, 205–223. https://doi.org/10.1023/A:1005592401947

Lord, F.M., Novick, M.R., Birnbaum, A., 1968. Statistical theories of mental test scores, Statistical theories of mental test scores. Addison-Wesley, Oxford, England.

Mallidi, A., Meza-Cervera, T., Kircanski, K., Stringaris, A., Brotman, M.A., Pine, D.S., Leibenluft, E., Linke, J.O., 2023. Robust caregiver-youth discrepancies in irritability ratings on the affective reactivity index: An investigation of its origins. J Affect Disord S0165-0327(23)00451–2. https://doi.org/10.1016/j.jad.2023.03.091

March, J.S., Parker, J.D.A., Sullivan, K., Stallings, P., Conners, C.K., 1997. The Multidimensional Anxiety Scale for Children (MASC): Factor Structure, Reliability, and Validity. Journal of the American Academy of Child & Adolescent Psychiatry 36, 554–565. https://doi.org/10.1097/00004583-199704000-00019

Mazefsky, C.A., Day, T.N., Siegel, M., White, S.W., Yu, L., Pilkonis, P.A., Autism and Developmental Disabilities Inpatient Research Collaborative (ADDIRC), 2018. Development of the Emotion Dysregulation Inventory: A PROMIS®ing Method for Creating Sensitive and Unbiased Questionnaires for Autism Spectrum Disorder. J Autism Dev Disord 48, 3736–3746. https://doi.org/10.1007/s10803-016-2907-1

Morris, A.C., Ibrahim, Z., Heslin, M., Moghraby, O.S., Stringaris, A., Grant, I.M., Zalewski, L., Pritchard, M., Stewart, R., Hotopf, M., Pickles, A., Dobson, R.J.B., Simonoff, E., Downs, J., 2023. Assessing the feasibility of a web-based outcome measurement system in child and adolescent mental health services – myHealthE a randomised controlled feasibility pilot study. Child and Adolescent Mental Health 28, 128–147. https://doi.org/10.1111/camh.12571

Myin-Germeys, I., Kasanova, Z., Vaessen, T., Vachon, H., Kirtley, O., Viechtbauer, W., Reininghaus, U., 2018. Experience sampling methodology in mental health research: new insights and technical developments. World Psychiatry 17, 123–132. https://doi.org/10.1002/wps.20513

NHS England » Future in Mind: Children and Young People's Mental Wellbeing [WWW Document], n.d. URL https://www.england.nhs.uk/blog/martin-mcshane-14/ (accessed 4.9.23).

Nicol, G., Wang, R., Graham, S., Dodd, S., Garbutt, J., 2022. Chatbot-Delivered Cognitive Behavioral Therapy in Adolescents With Depression and Anxiety During the COVID-19 Pandemic: Feasibility and Acceptability Study. JMIR Formative Research 6, e40242. https://doi.org/10.2196/40242

Nunnally, J., Bernstein, I., 1994. Psychometric Theory, 3rd ed. McGraw-Hill.

Poznanski, E.O., Grossman, J.A., Buchsbaum, Y., Banegas, M., Freeman, L., Gibbons, R., 1984. Preliminary Studies of the Reliability and Validity of the Children's Depression Rating Scale. Journal of the American Academy of Child Psychiatry 23, 191–197. https://doi.org/10.1097/00004583-198403000-00011

Reisner, S.L., Katz-Wise, S.L., Gordon, A.R., Corliss, H.L., Austin, S.B., 2016. Social epidemiology of depression and anxiety by gender identity. J Adolesc Health 59, 203–208. https://doi.org/10.1016/j.jadohealth.2016.04.006

Reynolds, W.M., 2004. The Reynolds Adolescent Depression Scale-Second Edition (RADS-2), in: Comprehensive Handbook of Psychological Assessment, Vol. 2: Personality Assessment. John Wiley & Sons, Inc., Hoboken, NJ, US, pp. 224–236.

Rognstad, K., Wentzel-Larsen, T., Neumer, S.-P., Kjøbli, J., 2023. A Systematic Review and Meta-Analysis of Measurement Feedback Systems in Treatment for Common Mental Health Disorders. Adm Policy Ment Health 50, 269–282. https://doi.org/10.1007/s10488-022-01236-9

Schwab-Stone, M.E., Shaffer, D., Dulcan, M.K., Jensen, P.S., Fisher, P., Bird, H.R., Goodman, S.H., Lahey, B.B., Lichtman, J.H., Canino, G., Rubio-Stipec, M., Rae, D.S., 1996. Criterion validity of the NIMH Diagnostic Interview Schedule for Children Version 2.3 (DISC-2.3). J Am Acad Child Adolesc Psychiatry 35, 878–888. https://doi.org/10.1097/00004583-199607000-00013

Scrutton, A.P., 2017. Epistemic Injustice and Mental Illness, in: The Routledge Handbook of Epistemic Injustice. Routledge.

Shaffer, D., Fisher, P., Dulcan, M.K., Davies, M., Piacentini, J., Schwab-Stone, M.E., Lahey, B.B., Bourdon, K., Jensen, P.S., Bird, H.R., Canino, G., Regier, D.A., 1996. The NIMH Diagnostic Interview Schedule for Children Version 2.3 (DISC-2.3): description, acceptability, prevalence rates, and performance in the MECA Study. Methods for the Epidemiology of Child and Adolescent Mental Disorders Study. J Am Acad Child Adolesc Psychiatry 35, 865–877. https://doi.org/10.1097/00004583-199607000-00012

Shaffer, D., Gould, M.S., Brasic, J., Ambrosini, P., Fisher, P., Bird, H., Aluwahlia, S., 1983. A Children's Global Assessment Scale (CGAS). Archives of General Psychiatry 40, 1228–1231. https://doi.org/10.1001/archpsyc.1983.01790100074010

Spence, S.H., Barrett, P.M., Turner, C.M., 2003. Psychometric properties of the Spence Children's Anxiety Scale with young adolescents. Journal of Anxiety Disorders 17, 605–625. https://doi.org/10.1016/S0887-6185(02)00236-0

Stevens, K., Ratcliffe, J., 2012. Measuring and valuing health benefits for economic evaluation in adolescence: an assessment of the practicality and validity of the child health utility 9D in the Australian adolescent population. Value Health 15, 1092–1099. https://doi.org/10.1016/j.jval.2012.07.011

Streiner, D.L., Norman, G.R., Cairney, J., n.d. Health Measurement Scales: A practical guide to their development and use. Oxford University Press.

Stringaris, A., 2021. Sources of normativity in childhood depression. Eur Child Adolesc Psychiatry 30, 1663–1665. https://doi.org/10.1007/s00787-021-01891-7

Stringaris, A., Goodman, R., Ferdinando, S., Razdan, V., Muhrer, E., Leibenluft, E., Brotman, M.A., 2012. The Affective Reactivity Index: a concise irritability scale for clinical and research settings. Journal of Child Psychology and Psychiatry 53, 1109–1117. https://doi.org/10.1111/j.1469-7610.2012.02561.x

Swanson, J., Deutsch, C., Cantwell, D., Posner, M., Kennedy, J.L., Barr, C.L., Moyzis, R., Schuck, S., Flodman, P., Spence, M.A., Wasdell, M., 2001. Genes and attention-deficit hyperactivity disorder. Clinical Neuroscience Research 1, 207–216. https://doi.org/10.1016/S1566-2772(01)00007-X

Urbina, S., 2016. Psychological Testing: An Overview, in: Encyclopedia of Mental Health. Academic Press, Elsevier.

Wakschlag, L.S., Hill, C., Carter, A.S., Danis, B., Egger, H.L., Keenan, K., Leventhal, B.L., Cicchetti, D., Maskowitz, K., Burns, J., Briggs-Gowan, M.J., 2008. Observational Assessment of Preschool Disruptive Behavior, Part I: reliability of the Disruptive Behavior Diagnostic Observation Schedule (DB-DOS). J Am Acad Child Adolesc Psychiatry 47, 622–631. https://doi.org/10.1097/CHI.0b013e31816c5bdb

Weisz, J.R., Chorpita, B.F., Frye, A., Ng, M.Y., Lau, N., Bearman, S.K., Ugueto, A.M., Langer, D.A., Hoagwood, K.E., Research Network on Youth Mental Health, 2011. Youth Top Problems: using idiographic, consumer-guided assessment to identify treatment needs and to track change during psychotherapy. J Consult Clin Psychol 79, 369–380. https://doi.org/10.1037/a0023307