Child Psychiatry & Human Development (2022) 53:737–753
https://doi.org/10.1007/s10578-021-01149-y

ORIGINAL ARTICLE

Are We Comparing Apples with Oranges? Assessing Improvement Across Symptoms, Functioning, and Goal Progress for Adolescent Anxiety and Depression

Karolin Rose Krause[1,2] · **Julian Edbrooke-Childs**[1,2] · **Rosie Singleton**[1,2] · **Miranda Wolpert**[1,3]

Accepted: 22 February 2021 / Published online: 7 April 2021

Abstract

Strategies for comparing routinely collected outcome data across services or systems include focusing on a common indicator (e.g., symptom change) or aggregating results from different measures or outcomes into a comparable core metric. The implications of either approach for judging treatment success are not fully understood. This study drew on naturalistic outcome data from 1641 adolescents with moderate or severe anxiety and/or depression symptoms who received routine specialist care across 60 mental health services in England. The study compared rates of meaningful improvement between the domains of internalizing symptoms, functioning, and progress towards self-defined goals. Consistent cross-domain improvement was observed in only 15.6% of cases. Close to one in four (24.0%) young people with reliably improved symptoms reported no reliable improvement in functioning. Inversely, one in three (34.8%) young people reported meaningful goal progress but no reliable symptom improvement. Monitoring systems that focus exclusively on symptom change risk over- or under-estimating actual impact, while aggregating different outcomes into a single metric can mask informative differences in the number and type of outcomes showing improvement. A move towards harmonized outcome measurement approaches across multiple domains is needed to ensure fair and meaningful comparisons.

Keywords Adolescents · Depression · Anxiety · Outcome · Functioning · Personalized measures

Introduction

Anxiety and depression are among the most common mental health conditions in young people worldwide [1–3]. In the absence of effective treatment, early-onset depression and anxiety can have significant adverse effects on mental health and socio-economic outcomes across the life course [4–7]. Treatments that have demonstrated efficacy in clinical trials [8–14] often do not yield the expected results in clinical practice [15, 16]. Routinely collected outcome data has an important role in enhancing the effectiveness of routine care for adolescent anxiety and depression, by informing adjustments to individual treatment plans, service planning, target setting and comparisons of effectiveness across settings and care models [17–20].

Services wishing to embed routine outcome monitoring face the challenge of having to identify the most important outcomes to measure, as well as the most suitable measurement instruments to track these outcomes. Notably, such instruments should be feasible, acceptable, valid, and reliable [21]. In addition, consultations with youth, families, and clinicians indicate that outcome measurement should be tailored to individual cases, and assess change holistically [22–24]. Services further face system-level challenges related to synthesizing and benchmarking data obtained from different measurement approaches. While standards for the routine measurement of outcomes for adolescent depression are beginning to emerge [21], they are not yet widely implemented.

One possible avenue for managing this challenge is to build reporting systems around a single outcome metric.

✉ Karolin Rose Krause
karolin.krause.16@ucl.ac.uk

[1] Research Department for Clinical, Educational and Health Psychology, University College London, Gower Street, Bloomsbury, London WC1E 6BT, UK

[2] Evidence-Based Practice Unit, Anna Freud National Centre for Children and Families, 4–8 Rodney Street, London N1 9JH, UK

[3] Wellcome Trust, 215 Euston Rd, Bloomsbury, London NW1 2BE, UK

For example, the US National Committee for Quality Assurance includes an indicator of symptomatic recovery derived from the Patient Health Questionnaire 9 (PHQ-9) [25] in its Healthcare Effectiveness Data and Information Set (HEDIS) to measure care quality for depression. Such approaches mirror common practice in clinical trials, where the designated primary outcome measure typically tracks symptom change [26]. While a single indicator provides the clarity that commissioners and policymakers require for decision-making and reporting purposes [27], there are questions about the extent to which a symptom score change by itself represents meaningful improvement in young people's lives, including in their daily functioning [28–32].

A second avenue is to measure outcomes more holistically, across multiple domains (and possibly by using multiple measures within the same outcome domain), and to aggregate results into a standardized composite metric to simplify reporting and benchmarking (e.g., [33]). For example, young people could be considered "reliably improved" if they demonstrated reliable improvement on at least one of the individual measures or outcome domains considered, and if there was no reliable decline on any other [34]. Two outcome domains for consideration alongside symptoms are functioning and progress towards self-defined goals. Measures of functioning provide insight into how symptoms impact on daily life, and can anchor and contextualize symptom scores [32]. Progress towards self-defined goals is measured via a personalized scale where item content is determined by individual service users, enabling a person-centred progress assessment [35, 36].

The implications of either approach for determining treatment effectiveness for individual cases, and at a service level, are not well understood. There is limited evidence about the extent to which measures that purport to capture the same construct converge in their ratings for individual cases, or the extent to which improvement in one outcome domain translates into improvement in another domain [37, 38]. Existing research suggests that symptom change often exceeds change in functioning [39–41], but tends to be inferior to subjective perceptions of change or progress towards self-defined goals, as measured by personalized instruments [42, 43].

Existing studies have tended to use cross-diagnostic samples, and the extent to which their findings apply to adolescent anxiety and depression is unclear. There is some evidence that externalizing disorders are associated with higher functional impairment than internalizing disorders [44, 45], and the association between changes in symptoms and functioning may vary across clinical presentations. In addition, only two studies have examined the convergence of change ratings across different outcome domains at an individual level, and both were limited by small samples of around 120 cases [40, 42]. One additional study examined overall rates of reliable change between measures of symptom change, functioning, and progress towards self-defined goals, but without examining the extent to which reliable change ratings for different outcome domains converged for individual service users.

The Present Study

Building on previous research, this study examined the convergence of meaningful improvement ratings [46] between (a) two measures of internalizing symptoms; (b) two measures of psychosocial functioning; and (c) between aggregate ratings in the domains of symptoms, functioning, and progress towards self-defined goals in a sample of adolescents aged 12–18 years with moderate or severe depression and/or anxiety symptoms who accessed routine specialist mental health care. Meaningful improvement was defined as reliable improvement on a standardized scale, and as meaningful improvement on an idiographic, goal-based outcome measure. The study used naturalistic data obtained through the routine administration of self-report measures that are widely used across a range of mental health settings in England [47].

Methods

General Setting

This study was a post-hoc analysis of naturalistic outcome data collected by specialist child and adolescent mental health services in England between 2011 and 2015, and collated by the Child Outcomes Research Consortium (CORC)[1] as part of NHS England's Children and Young People's Improving Access to Psychological Therapies (CYP IAPT) service transformation initiative [34, 48]. Services could select from several standardized youth- and/or parent-reported measures of symptoms and functioning, as well as a personalized measure of progress towards self-defined goals [49].

Participants and Process

The full CYP IAPT dataset included 96,325 case records of which 23,373 cases attended at least one appointment following the initial assessment [34]. For inclusion in this analysis, cases had to be aged 12–18 years and have moderate to severe anxiety or depression symptoms as indicated by clinicians on the Current View tool [50]. This was true for 15,352 case records. As this study examined rates of

[1] CORC is a professional learning collaboration.

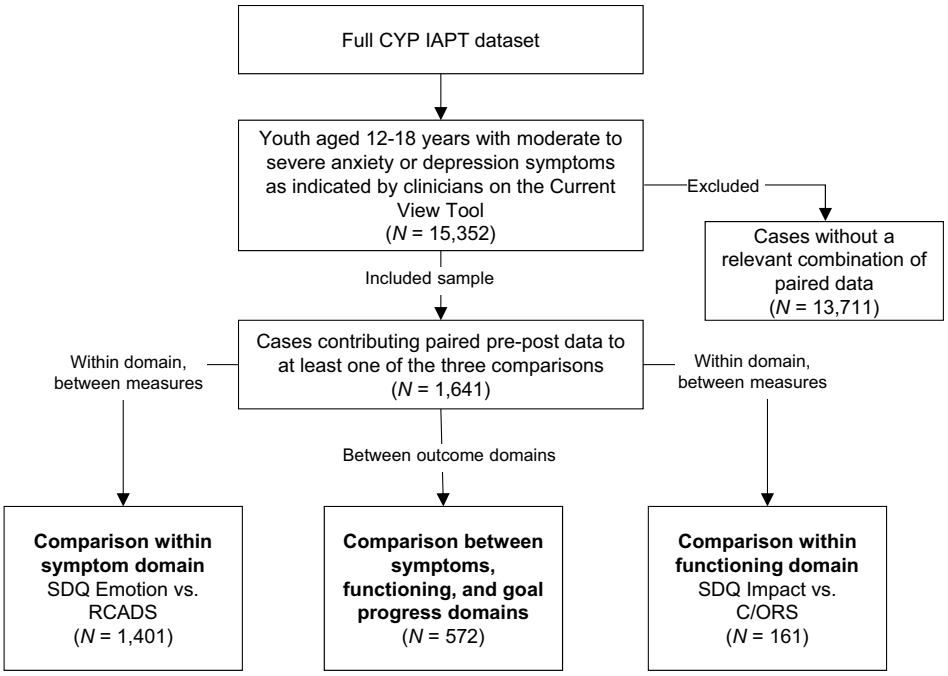

Fig. 1 Flowchart of the analytical process and sampling

Note: CYP IAPT = Children and Young People's Improving Access to Psychological Therapies.

meaningful improvement, cases rated as showing only mild anxiety or depression symptoms at first assessment were not included, to ensure sufficient scope for improvement.

Cases also needed to have *paired data* on a relevant combination of outcome measures, that is, they needed to have completed the target measures on at least two occasions. Cases were required to either have paired data on both internalizing symptom measures ($N = 1401$), or on both functioning measures ($N = 161$); or on at least one symptom measure, one functioning measure, and the goal-based outcome measure ($N = 572$) to be included. This formed an overall analytical sample of 1641 cases. Because services were not mandated to use specific measures, different tools were administered in a variety of combinations, leading to a high amount of missing data on some measures. The breakdown of the study sample into the three analytical subsamples is illustrated in Fig. 1.

Of the 1641 cases included, 75.1% were female, 87.4% identified as white British, and the mean age was 14.8 years (SD = 1.47). On the Current View, 49.5 and 22.1% of the included sample were rated as showing moderate or severe anxiety symptoms, respectively; and 55.6 and 11.6%, respectively, were rated as showing moderate or severe depression symptoms (see Table 1). Data originated from 60 specialist child and adolescent mental health services and the average contact length was 31.0 weeks. Compared with cases that were excluded due to missing data, included cases were more likely to be female, $\chi^2(1) = 79.05$, $p < 0.001$; and to identify as white, $\chi^2(1) = 23.62$, $p < 0.001$. Youth in the included sample were slightly more likely to be classified as severely anxious $\chi^2(1) = 28.27$, $p < 0.001$, or severely depressed, $\chi^2(1) = 21.31$, $p < 0.001$. There was no difference in mean age between the included and excluded samples, $t(15{,}350) = -0.83$, $p = 0.41$.

For 24.6% of the included sample, clinician ratings on the Current View indicated at least moderate self-harm alongside symptoms of anxiety or depression. In addition, 33.5% of youth were rated as having moderate to severe co-occurring difficulties in at least one other problem area, with symptoms of post-traumatic stress disorder (13.0%) and obsessive–compulsive disorder (12.4%) being the most common co-occurring presenting problems (see Table 1). The naturalistic dataset does not provide information on the primary diagnosis. As such, it was not possible to establish whether youth with co-occurring problems were treated primarily for anxiety or depression. Similarly, the Current View tool does not capture whether moderate or severe symptoms in any given problem area were accompanied by a formal clinical diagnosis.

Young people received specialist routine care as provided by the 60 child and adolescent mental health services that contributed data to this study. Treatment approaches were

Table 1 Characteristics of the included and excluded samples

Characteristics	Excluded sample	Included sample			
		Total included sample	Symptom comparison	Functioning comparison	Domain comparison
	N = 13,711 (%)	*N* = 1641 (%)	*N* = 1401 (%)	*N* = 161 (%)	*N* = 572 (%)
Sex (% female)	69.3	75.1	74.9	73.3	75.7
Ethnicity (% White British)	84.4	87.4	87.5	85.0	86.2
Current view anxiety rating					
% moderate anxiety	49.7	49.5	49.3	41.0	51.7
% severe anxiety	20.5	22.1	21.8	26.3	22.4
Current view depression rating					
% moderately depressed	54.93	55.6	55.2	63.8	55.4
% severely depressed	10.0	11.6	12.32	11.3	9.1
Current view ratings for co-occurring problems (moderate or severe)					
% self-harm	24.6	26.1	26.8	25.0	25.7
% PTSD	17.1	13.0	13.3	9.5	11.0
% OCD	12.1	12.4	12.1	8.8	13.6
% CD or ODD	11.3	6.0	6.1	6.8	4.6
% eating disorder	10.1	9.0	9.1	8.9	5.8
% ADHD/hyperactivity	8.6	5.7	5.5	6.0	6.0
% psychosis or bipolar disorder	8.3	5.6	5.5	6.2	4.8
% substance use	4.0	1.6	1.7	4.1	1.3
	M (SD)	*M* (SD)	*M* (SD)	*M* (SD)	*M* (SD)
Age (in years)	14.7 (1.54)	14.8 (1.47)	14.8 (1.47)	14.9 (1.36)	14.8 (1.46)
Mean contact length (weeks)[a]	N/A	31.0 (20.39)	31.0% (20.63)	36.7 (21.46)	31.2 (20.02)
No. of services	77	60	60	32	53

ADHD attention deficit hyperactivity disorder, *CD* conduct disorder, *OCD* obsessive–compulsive disorder, *ODD* oppositional defiant disorder, *PTSD* post-traumatic stress disorder

[a]Length of contact was computed based on the dates of the very first and very last assessment completed on the RCADS, the SDQ or the C/ORS

not consistent across the sample, but varied according to the modalities and protocols used by each service, and the needs of individual cases. Data on the type of treatment received was available for 50.2% of the included sample. The most common type of therapy received by youth was cognitive behavioural therapy (65.3%). Other treatments included systemic family therapy (16.6%), psychodynamic or psychoanalytic psychotherapy (10.6%), brief solution-focused therapy (9.2%) and interpersonal psychotherapy (8.1%).

Ethical Review

As this study was a secondary analysis of routinely collected data, ethical review was not required [51].

Measures

Current View [50]

The current view is a screening tool to be completed by clinicians at first contact to provide a snapshot view of a case profile. The tool captures presenting problems, complexity factors and contextual problems, but not formal diagnoses. Clinicians are instructed to complete the Current View by drawing on all relevant information they have available at the time of tool completion. This includes information obtained through conversations during intake and initial assessment, information shared in the referral process (e.g., by other health professionals, teachers, or social workers), and scores from self- or clinician-reported outcome measures [50]. Presenting problems are assessed via 30 problem descriptions that map onto ICD-11 diagnostic criteria relevant to children and adolescents [52]. Based on the information available to them, clinicians rate the perceived severity of distress and impairment for each problem on a scale from 0 (none) to 3 (severe). Anxiety is identified via six problem descriptions that respectively indicate symptoms of separation anxiety, social phobia, generalized anxiety, panic disorder, agoraphobia, and specific phobia. Depression is identified through problem description number 9: "Depression/low mood (Depression)." As the naturalistic dataset did not contain information about the formal diagnoses assigned by the treating clinicians, the Current View problem descriptions were used as a proxy in this study to appraise the baseline severity of anxiety, depression, and co-occurring difficulties.

Revised Children's Anxiety and Depression Scale (RCADS)—Total Score [53]

The RCADS is a 47-item self-report scale for 8–18-year-olds, measuring the frequency of symptoms associated with depression and anxiety. Young people are asked to state how frequently they experience each symptom, using a four-point Likert scale ranging from 0 (never) to 3 (always). A recall period is not specified. The RCADS consists of six subscales measuring symptoms of major depressive disorder (ten items), generalized anxiety disorder (six items), separation anxiety disorder (seven items), social phobia (nine items), panic disorder (nine items), and obsessive–compulsive disorder (six items), in line with DSM-IV dimensions [54]. Subscale scores can be summed to compute overall anxiety, depression, and internalizing symptom scores. The RCADS has demonstrated good internal consistency, test–retest reliability, and construct validity [53, 55–58].

Strengths and Difficulties Questionnaire (SDQ)—Emotional Problems Subscale [59, 60]

The SDQ is a 25-item self-report measure of psychosocial difficulties in children and adolescents aged 4–16 years. Respondents are asked to rate problem descriptions on a 3-point scale, from 0 ("not true") to 2 ("certainly true"). An assessment version of the SDQ to be used at first measurement asks about psychosocial problems with reference to the past six months. A follow-up version of the SDQ, to be used at subsequent measurement time points, enquires about psychosocial difficulties in the past month. The SDQ includes a five-item emotional problems subscale that captures unhappiness, worries, clinginess, fears, and somatic symptoms. The five items can be summed to obtain a total subscale score ranging from 0 to 10. While the SDQ as a whole has been widely used and validated with regards to its internal consistency, test–retest reliability, construct validity and predictive validity [61–64], internal consistency for the emotional symptoms subscale has been shown to be questionable, with a Cronbach's Alpha of 0.66 [62].

Strengths and Difficulties Questionnaire (SDQ)—Impact Supplement [65]

The SDQ Impact Supplement assesses the impact of psychosocial difficulties captured by the regular SDQ on a young person's daily life. It probes into the duration and degree of distress, and into the impact on home life, friendships, classroom learning and leisure activities. The assessment version of the SDQ Impact enquires about the impact of psychosocial difficulties overall, while the follow-up version enquires about the impact of psychosocial difficulties during the past month. The five items on distress and impact are scored on a 3-point scale from 0 ("not at all/only a little"), to 2 ("a great deal") and summed to compute a total score ranging from 0 to 10. The measure's developers report good internal consistency, with a Cronbach's Alpha of 0.81 [62].

Table 2 Parameters used to determine the RCI for each standardized measure in the study sample

Measure	N	M_{T1} (SD)	M_{T1}–M_{T2} (SD)	Cronbach alpha	RCI/MCI threshold
SDQ emotion	1577	7.13 (2.16)	− 1.53 (2.62)	0.64	3.62
RCADS	1427	69.85 (25.30)	− 18.45 (26.91)	0.95	15.68
SDQ impact	636	4.32 (2.43)	− 1.71 (2.79)	0.65	3.97
C/ORS	198	20.50 (8.45)	6.59 (9.55)	0.87	8.40

Child Outcome Rating Scale (CORS) and Outcome Rating Scale (ORS)—Total Score [66]

The CORS and the ORS are four-item self-report measures of general distress and psychosocial functioning. The CORS was designed for use with children aged 6–12 years and is more child-friendly in wording and layout than the ORS, which was designed for youth aged 13 and older [67]. Both versions consist of four items that cover personal well-being, interpersonal functioning (e.g., with regards to family and close relationships), social functioning (e.g., at school or work), and overall well-being. The CORS asks about how things are going in general, while the ORS enquires about how things have been over the past week. Responses to each question are recorded as markings on a 10 cm visual analogue scale. Scoring is done by measuring the length between the starting point of the visual analogue scale and the marker, and by converting the distance from centimetres into score points (i.e., ten is the highest-possible score). A total score is computed by summing the four subscale scores. Validation studies have reported good internal consistency (between $\alpha = 0.81$ and 0.97) [68–70], but mixed findings for test–retest reliability ($r = 0.66$–0.81) [66, 68]. For this analysis, CORS and ORS scores were combined into a composite score.

Goals Based Outcomes (GBO) [71]

The GBO tool is a personalized outcome measure, designed primarily with clinical utility in mind [72]. Young people can define a number of treatment goals, the top three of which are typically used for outcomes reporting [73]. Progress is rated periodically on a scale from 0 ("goal not at all met") to 10 ("goal reached"), with young people indicating how they would rate their progress on the given day [74]. Only goals defined primarily by adolescents themselves were considered for this analysis. Data on the reliability of the GBO is not currently available [75].

Statistical Analysis

Assessing Meaningful Change

The criterion used in this study to determine the salience of individual-level change was the reliable change index (RCI) [76] for standardized measures, and a meaningful change index for the GBO. The RCI determines the amount of change required to demarcate improvement beyond fluctuations attributable to measurement error [76, 77]. The RCI is calculated by dividing the difference between T1 and T2 scores by the standard error of the difference between the two measurements (see Online Appendix for details). We computed the RCI based on the standard deviation of the mean T1 score and the measure's internal consistency at T1 in the subsample that contributed paired data on the relevant measure to any of the three comparisons. The SDQ Emotion and SDQ Impact demonstrated questionable internal consistency ($\alpha = 0.64$ and 0.65, respectively). Internal consistency was good on the C/ORS ($\alpha = 0.87$) and excellent on the RCADS ($\alpha = 0.95$). To be considered as reliably improved on a given measure, individuals had to demonstrate a pre-post score difference exceeding RCI thresholds of 15.68 for the RCADS, 3.62 for the SDQ Emotions, 3.97 on the SDQ Impact, and 8.40 on the C/ORS (Table 2).

The reliable change index for the GBO has previously been defined as a movement by at least 2.45 along the goal progress scale, where progress scores are aggregated across the three goals [43]. However, as service users were free to define fewer than three goals in CYP IAPT, the incidence of missing data was high. We chose to compute an alternative meaningful change index that would use all available GBO data, without requiring complete measurements on all three goals. Meaningful change on the GBO was defined as youth showing an improvement by at least three scale points on any completed goal, without equivalent deterioration on any other available goal [33]. For the sake of brevity, we will use the shorthand term "improved" to describe both reliable improvement and meaningful improvement.

We conducted a series of bivariate comparisons of improvement between the SDQ Emotion and the RCADS within the symptom domain ($N = 1401$); between the SDQ Impact and the C/ORS within the functioning domain ($N = 161$); and between pairs of outcome domains ($N = 572$). We also conducted a multivariate comparison of improvement rates between the three outcome domains of symptoms, functioning, and goal progress. Composite improvement indices were computed for the symptom and functioning domains by defining as reliably improved those who showed reliable improvement on at least one of the two measures

within each domain, and no reliable deterioration on the other (see Table A2 in the Online Appendix).

Since few cases showed reliable or meaningful deterioration (ranging from 1.9 to 7.0%, see Table A1 in the Online Appendix), cross tables of dichotomized improvement ratings were computed, distinguishing only between improvement versus no improvement (including deterioration) to avoid cell sample sizes below the reportable minimum [78]. We computed McNemar's test of correlated proportions [79] to assess the likelihood of no agreement, and Cohen's Kappa for chance-corrected agreement to estimate the level of agreement in bivariate comparisons [80]. Fleiss' kappa was computed to assess agreement across all three domains [81].

Assessing Discrepancies in Assessment Timelines for the Standardized Measures

As this study used naturalistic data, we expected outcome assessment timelines to vary between individuals, as well as within individual cases with regard to different measurement instruments. We conducted descriptive analysis to examine the mean and median time between the first assessment (hereafter "T1") and last assessment (hereafter "T2") for each standardized measure. A Wilcoxon signed rank test was conducted to assess the significance of any observed differences in median time for the pairs of symptom and functioning measures, respectively. Univariate logistic regression was conducted to examine whether chances of reliable improvement increased with increasing time elapsed between T1 and T2. Data on assessment timelines for the GBO were not available.

Results

Comparing Improvement Within the Symptom and Functioning Domains

Figures illustrating the results below are included in the Online Appendix (page 4).

Convergence of Improvement Ratings Between Symptom Measures

In the sample with paired data on both internalizing symptom measures ($N=1401$), reliable improvement rates were considerably higher on the RCADS (49.3%) than on the SDQ Emotion (22.9%, see Table 3). At an individual level, improvement ratings were discrepant in close to one third of cases. Of all cases who did not improve on the SDQ Emotion, more than a third (37.4%) did improve on the RCADS. Of the cases that improved on the RCADS, 58.5% failed to improve on the SDQ Emotion. McNemar's test of correlated proportions showed a significant difference in improvement ratings [$\chi^2(1)=312.6$; $p<0.001$], and Cohen's kappa indicated fair agreement ($\kappa=0.37$; $p<0.001$) between the two measures. Only 20.5% of cases improved on both symptom measures.

Convergence of Improvement Ratings Between Functioning Measures

In the sample with paired data on both functioning measures ($N=161$), reliable improvement was considerably higher on the C/ORS (39.8%) than on the SDQ Impact (28.0%, Table 3). At an individual level, improvement ratings were discrepant in 42.9% of cases. Of all cases improving on the C/ORS, over two thirds (68.8%) showed no improvement on the SDQ Impact. At the same time, of the cases that improved on the SDQ Impact, 55.6% did not improve on the C/ORS (see Fig. 2). McNemar's test of correlated proportions [$\chi^2(1)=5.23$, $p=0.030$] indicated marginally significant disagreement. Cohen's kappa ($\kappa=0.06$, $p=0.224$) was not statistically significant, likely due to the small sample size. Only 12.4% of cases improved on both functioning measures.

Comparing Improvement Across Domains

In the sample with paired data on at least one symptom measure, one functioning measure, and the goal-based outcome measure ($N=527$), 69.9% of adolescents meaningfully improved their goal progress, 43.0% improved their internalizing symptoms, and 28.9% improved their functioning (see Table A1 in the Online Appendix). Figures illustrating these results are included in the Online Appendix (pages 4–5).

Comparing Improvement Between the Symptom and Functioning Domains

One third (33.7%) of cases showed discrepant improvement ratings across the symptom and functioning domains. Of all cases that improved in the symptom domain, 55.7% did not improve their functioning (Table 3). In turn, of those improving their functioning, 33.9% did not improve their symptoms. McNemar's test showed a significant difference in improvement ratings between the symptom and functioning domains [$\chi^2(1)=33.99$; $p<0.001$]. Cohen's kappa indicated fair agreement ($\kappa=0.28$; $p<0.001$). Only 19.1% of cases improved in both domains.

Comparing Improvement Between the Symptom and Goal Progress Domains

Around 42.7% of cases showed discrepant improvement ratings across the symptom and goal domains. Of all

Table 3 Disagreement between measures and domains (bivariate comparisons)

First measure/domain		Second measure/domain		
		Not improved	Improved	Total
		n (%)	*n* (%)	*n* (%)
Within the symptom domain				
		RCADS		
SDQ emotion	Not improved	676 (48.3)	404 (28.8)	1,080 (77.1)
	Improved	34 (2.4)	287 (20.5)	321 (22.9)
	Total	710 (50.7)	691 (49.3)	1401 (100)
Within the functioning domain				
		C/ORS		
SDQ impact	Not improved	72 (44.7)	44 (27.3)	116 (72.0)
	Improved	25 (15.5)	20 (12.4)	45 (28.0)
	Total	97 (60.2)	64 (39.8)	161 (100)
Across the symptom, functioning, and goal progress domains (paired comparisons)				
		Functioning		
Internalizing symptoms	Not improved	270 (47.2)	56 (9.8)	326 (57.0)
	Improved	137 (24.0)	109 (19.1)	246 (43.0)
	Total	407 (71.2)	165 (28.8)	572 (100)
		Goal progress		
Internalizing symptoms	Not improved	127 (22.2)	199 (34.8)	326 (57.0)
	Improved	45 (7.9)	201 (35.1)	246 (43.0)
	Total	172 (30.1)	400 (69.9)	572 (100)
		Goal progress		
Functioning	Not improved	139 (24.3)	268 (46.9)	407 (71.2)
	Improved	33 (5.8)	132 (23.1)	165 (28.8)
	Total	172 (30.1)	400 (69.9)	572 (100)

those who improved their goal progress, 49.8% showed no improvement in internalizing symptoms, while of those not improving on the GBO, 26.2% still improved their symptoms. McNemar's test of correlated proportions showed a significant difference in improvement ratings between the symptom and goal progress domains [$\chi^2(1) = 97.20$; $p < 0.001$] while Cohen's kappa indicated slight agreement ($\kappa = 0.19$; $p < 0.001$). A third of cases (35.1%) showed improvement in both domains.

Comparing Improvement Between the Functioning and Goal Progress Domains

More than half of cases (52.6%) showed discrepant improvement ratings across the functioning and goal domains. Of all cases improving their goal progress, 67.0% failed to improve their functioning. In turn, of those who did improve their functioning, 20.0% showed no improvement in goal progress. McNemar's test of correlated proportions showed a significant difference in improvement ratings between the functioning and goal progress domains [$\chi^2(1) = 183.47$; $p < 0.001$]. Cohen's kappa ($\kappa = 0.10$; $p < 0.001$) indicated slight agreement. Only 23.1% of cases improved in both domains.

Comparing Improvement Between the Symptom, Functioning and Goal Progress Domains

When comparing rates of meaningful improvement across all three domains, 64.5% of cases showed discrepant ratings. Of all 572 cases considered, 20.0% did not show improvement in any domain, 27.3% improved exclusively in the goal progress domain, 19.6% improved their symptoms and goal progress but not their functioning, and 15.6% improved across all three domains [with 2–8% of cases showing other

Table 4 Disagreement between symptoms, functioning, and goal progress

	Functioning	Goal progress	
		Not improved	Improved
Internalizing symptoms			
Not improved	Not improved	114 (19.9%)	156 (27.3)
	Improved	13 (2.3%)	43 (7.5%)
Improved	Not improved	25 (4.4%)	112 (19.6%)
	Improved	20 (3.5%)	89 (15.6)

$N = 572$ (100%)

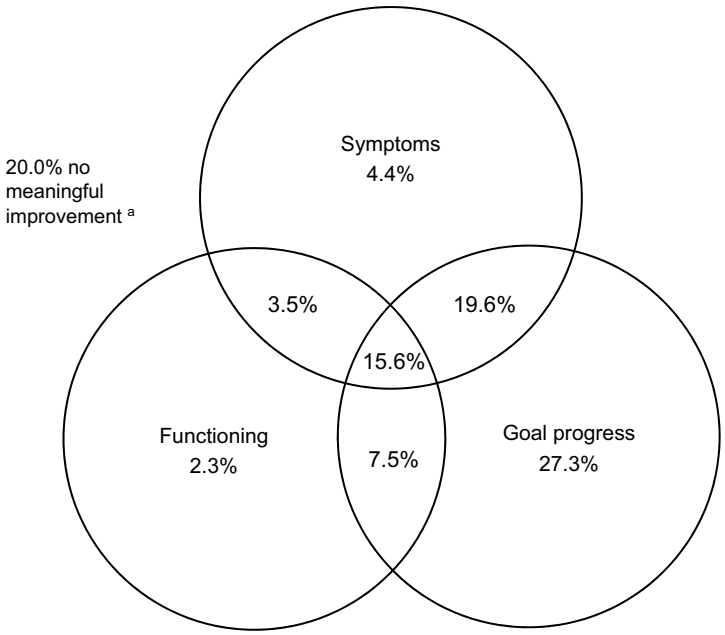

Fig. 2 Venn diagram of meaningful improvement.[a] Across all three outcome domains

Note. $N = 572$. Figures within the circles indicate percentages of reliable or meaningful improvement. The circles are not proportionate in size to the level of improvement observed in each domain.

[a] Meaningful improvement was determined based on the reliable change index for standardized measures, and based on a meaningful change index for the idiographic goal-based outcome measure.

combinations of change (Table 4, Fig. 2)]. Fleiss' kappa showed only slight agreement in improvement ratings across the three domains ($\kappa_f = 0.14$; $p < 0.001$).

Assessment Timelines for the Standardized Measures

Within the symptom measure comparison sample ($N = 1401$), the mean amount of time elapsed between T1 and T2 was 29.3 weeks for the RCADS (SD = 20.1; median = 25 weeks), and 28.6 weeks for the SDQ Emotions (SD = 19.4; median = 24.9 weeks). A Wilcoxon signed rank test showed no significant difference between the two measures ($Z = 0.433$, $p = 0.67$). For 78.4% of the sample, the number of weeks elapsed was identical on both measures (see Online Appendix). Within the functioning measure comparison sample ($N = 161$), the mean amount of time elapsed between T1 and T2 was 28.8 weeks (SD = 20.1; median = 24.9 weeks) for the SDQ Impact, and 20.65 weeks (SD = 17.3; median = 16.0 weeks) for the C/ORS. A Wilcoxon signed rank test showed a significant difference between the two measures ($Z = -5.261$, $p < 0.001$). The amount of time elapsed was identical across the two measures for only 14.9% of the sample. Given the longer average assessment period for the SDQ Impact, it might be expected that this measure captured higher rates of reliable improvement. However, a univariate logistic regression showed no significant association between increasing length of time between assessments and the odds of achieving reliable improvement on the RCADS, the SDQ Emotions or the C/ORS (see Online Appendix), and only a marginally significant association for the SDQ Impact (OR 1.02; $p = 0.046$; 95% CI 1.00–1.03).

Discussion

This study assessed levels of meaningful improvement among adolescents with moderate or severe anxiety and/or depression symptoms across five commonly used self-report measures of internalizing symptoms, functioning, and progress towards self-defined goals; and examined the convergence of improvement ratings within and between these three outcome domains. We found considerable disagreement between measures and domains. The two symptom measures yielded discordant ratings for close to one third of cases, and the two functioning measures for over 40%. Similar levels of bivariate discordance were observed between the three domains of symptoms, functioning, and goal progress. There were discrepancies in 64.5% of cases when considering all three domains simultaneously. Improvement rates were highest in the goal progress domain (69.9%), and lowest in functioning (28.9%). Improvement was observed consistently across all three domains for only 15.6% of cases.

Within the symptom domain, the RCADS showed higher levels of reliable improvement (49.3%) than the SDQ Emotion (22.9%), although there was no significant difference in average assessment timelines between the two measures. Both measures track internalizing symptoms, but the RCADS provides a more detailed assessment of symptoms related to major depression (via 10 items), and to five anxiety disorders (via 6–9 items for each disorder). The SDQ Emotion consists of just one item measuring low mood; three items capturing fears, worries, and clinginess; and one item capturing somatic symptoms. Our finding is in line with previous research suggesting that more broadly defined measures are less likely to capture treatment effects than more specific measures [82], which may need to be taken into account when choosing measures for clinical or research use [47, 72, 83].

Within the domain of functioning, the C/ORS showed higher levels of reliable improvement than the SDQ Impact (39.8 and 28.0%, respectively). Both measures are of comparable length and cover psychosocial functioning in the family, peer, and school context. Both also enquire about global notions of distress or well-being. But while the SDQ Impact probes into functional impairment caused by mental health problems, the C/ORS asks young people how well they were generally doing. Although functioning measures probing about disorder-specific impairment have been described as more sensitive to change than generic measures [84], the problem-specific SDQ Impact displayed lower levels of change than the generic C/ORS. This was in spite of the average period between the T1 and T2 assessments being longer for the SDQ Impact than for the C/ORS, which would have allowed more time for functional improvements to become apparent. This discrepancy may be due to differences in response scales. The SDQ Impact uses a four-point Likert scale, while the C/ORS uses a ten-point continuous scale that may be better able to capture nuanced change. The observed discrepancy may also be driven by the differential internal consistencies of both measures, which led to a more conservative reliable change threshold for the SDQ Impact.

Our finding of limited convergence between reliable improvement in symptoms and functioning, and of less observed change in functioning is consistent with previous research [39, 40]. Associations between measures of depression symptoms and functioning may vary based on the types of symptoms assessed, as some symptoms may explain a larger variation in functional impairment than others [85]. In addition, several studies in adult populations have found change in social and global functioning to lag behind change in depressive symptoms [41]. Symptom change may be an early sign of treatment response, while functioning may be slower to manifest but could indicate deeper or more

wide-reaching impact. Another possible reason for the lower rates of reliable improvement observed in functioning may be that functioning measures for young people have received less attention from psychometricians, compared with symptom measures. They tend to have weaker psychometric properties, and possibly weaker sensitivity to change [39, 41, 86]. Another possible explanation is that 83% of cases considered for the cross-domain comparison had paired data on the SDQ Impact but not on the C/ORS. The comparatively low rate of reliable improvement in functioning may therefore be driven by the comparatively low rates of improvement on this specific measure, which further highlights the importance of careful measure selection.

Meaningful improvement in goal progress was considerably higher than reliable improvement in internalizing symptoms and functioning. This was consistent with two previous studies comparing change between standardized and personalized measures for children and young people [42, 43]. It is also consistent with a number of studies evidencing the superior sensitivity to change of personalized measures in adult mental health [87–89]. While parent-reported GBO scores have been shown to correlate moderately with parent-reported SDQ total difficulty scores ($r = 0.3$–0.4) and clinician-rated functioning ($r = 0.4$–0.5), no significant correlation has yet been found for child-reported measures of goal progress and symptoms [43, 90]. The GBO may capture changes that are uniquely different from those assessed by standardized measures of symptoms and functioning. For reasons related to the structure of the naturalistic dataset, it was not possible to consider the qualitative content of goals defined by adolescents in this study. However, a previous study examining the content of young people's self-defined goals found that they covered themes such as independence, confidence, self-reflection, communicating feelings, and understanding anger, which are not covered by commonly used standardized measures of symptoms or functioning [91]. Similarly, a recent study of drop-out in a treatment trial for adolescent depression reported that some young people ended treatment within three sessions, because they felt they had achieved their personal treatment goals, although this was not reflected by standardized outcome measures [92].

Administrative and Clinical Implications

The outcome domains services choose to monitor, and the tools they select for this purpose have a bearing on who is judged to have realized a "good" outcome, and what service models warrant funding. Measurement tools purported to measure a similar concept cannot be assumed to be comparable even when a standardized indicator such as the RCI is computed to facilitate comparisons. Due to differential psychometric properties (e.g., reliability and sensitivity to change) and item content, apples may be compared with oranges. This can have critical real-world implications where these discrepancies are interpreted as true differences in clinician or service performance.

Our findings corroborate existing evidence that symptom change by itself does not represent a sufficient proximal indicator for overall treatment effectiveness, as it may over- or under-estimate change in other outcome domains. Our findings call for more multidimensional approaches to routine outcome measurement, but multidimensional approaches raise challenges for interpreting and reconciling conflicting results that are comparable to the challenges posed by combining data from multiple reporters [93]. While the aggregation of reliable or meaningful improvement rates across measures and outcome domains provides a means of simplification and reduces the risk of missing change where it does occur, it fails to discriminate between cases that improve across all measures and/or domains, and cases that improve in only one. Such approaches may mask nuances that could help distinguish usual from best practice.

Given the lack of comparability between different outcome measures and outcome domains, mental health systems and services should seek to (a) track several meaningful outcome domains to gain a more holistic picture of the changes achieved, and (b) consider applying standards for outcome measurement, which are beginning to emerge. Our findings demonstrate that a common, harmonized approach to outcome measurement is needed if outcomes are to be compared fairly across mental health services and settings. A global standard set of outcomes for child and youth anxiety and depression has recently been developed under the lead of the International Consortium for Health Outcomes Measurement (ICHOM) [21]. The standard set recommends tracking change in the domains of symptoms, functioning, and suicidal thoughts and behaviour, as a minimum, when providing routine care for youth with anxiety or depression. The standard set further recommends a suite of seven feasible, valid, and reliable measurement instruments (including a short version of the RCADS), and suggests a timeline for outcome measurement.

The high rate of meaningful improvement on the GBO indicates that many young people experience change that is not reflected by standardized symptom and functioning measures. Based on this finding, services should consider administering personalized measures alongside standardized ones, so as not to miss idiographic treatment impact. Goal setting has demonstrated clinical value, in addition to being a flexible means of progress tracking, in terms of improving retention in treatment [94], and strengthening adolescents' perceived self-awareness and problem-solving ability [95].

Future Research

Further research is needed that examines the sensitivity to change of commonly used outcome measures, and that compares the magnitudes of change that each instrument can reliably detect at an individual level. There are no established cut-off criteria for establishing when a scale is sufficiently sensitive to change [21, 96]. The International Society for Quality of Life Research (ISOQOL) recommends that a measure "should have evidence of responsiveness, including empirical evidence of changes in scores consistent with pre-defined hypotheses" (p. 1901) [97]. One promising avenue for future research is to explore the *minimally important difference*, that is the minimum magnitude of change that is perceived as meaningful by service users and their families on a given scale [96, 98]. Reporting rates of minimally important difference that are anchored in stakeholder perceptions alongside rates of reliable improvement that are statistically derived could help with interpreting and contextualizing change metrics.

It is currently not well understood why treatment effectiveness is more difficult to evidence in functioning, compared with symptom change [39, 86, 99]. More research is needed to assess the validity, reliability, and sensitivity to change of youth-reported functioning measures, to understand whether brief scales such as the SDQ Impact and C/ORS provide the best possible avenue for tracking treatment response in clinical practice, or whether more granular measures are needed. There would be value in comparing change trajectories for symptoms and functioning over time, to explore whether outcomes in both domains are likely to converge at certain time points or for specific subgroups.

Personalized measures may show a higher sensitivity to change, because they are tailored to capturing change in the problems most salient to service users, and that treatment should ideally focus on [71, 73, 100]. However, the high levels of change measured by the GBO might also stem from services defining goals that are "too easy" to achieve [100]. More research involving children, adolescents, families, and clinicians, is needed to confirm that the currently used threshold for meaningful change is appropriate [75]. In addition, future research should examine the GBO's sensitivity to change and convergence with other measures in relation to specific goal themes, and to ascertain the psychometric properties of goals and other personalized measures.

Limitations

The above-mentioned findings should be considered in the context of several limitations. First, the dataset analysed for this study has been described elsewhere as an example of naturalistic data that are *flawed*, *uncertain*, *proximate,* and *sparse* ("FUPS") [83]. There was a high incidence of missing data, which could not be explained by the observed variables, hence limiting possibilities for data imputation. Second, the two assessment time points included in this dataset provide a snapshot view of the change achieved. Due to the high number of cases lost to follow up, change trajectories could not be examined in detail across several time points. In addition, there was considerable variation in assessment time points for different measures, especially for the two functioning measures, which may have added to the lack of convergence observed. Third, as outlined above, data on the qualitative content of the goals captured through the GBO were not available. As such, it was not possible to examine the extent to which self-defined goals covered similar content as the standardized measures examined in this study. Similarly, no information was available about the extent to which these goals were clinically relevant and achievable within the time course of specialist mental health support. As such, it was not possible to assess how goal characteristics may have influenced the high rates of meaningful improvement reported on the GBO in this study.

For the comparison across domains, composite reliable improvement indices were computed that pulled available data from the four symptom and functioning measures. As suggested above, the differential reliability of these measures meant that different magnitudes of change were required by each tool to exceed the reliable change threshold. Reliable change ratings that informed the composite metric for functioning were pulled exclusively from the SDQ Impact for 83% of the cases considered (see Table A2, in the Online Appendix). Given that the SDQ Impact indicated only about half the amount of reliable change as the C/ORS in the within-domain comparison, the cross-domain comparisons might have shown less discrepant results (due to higher rates of improvement in the functioning domain), had a larger share of functioning ratings been informed by the C/ORS.

Another limitation of this study is that we could not consider levels of convergence in relation to deterioration ratings, as the number of cases showing reliable or meaningful deterioration on the relevant measures were too small. Distinguishing not just between improvers and non-improvers, but also between those not showing any change and those showing deterioration might have led to even higher levels of discrepancy between measures and outcome domains, as differential sensitivity to change might have led to diverging rates of deterioration as well as improvement.

Conclusions

Routinely collected outcomes data is a crucial tool for strengthening service effectiveness. In countries where routine outcome monitoring is well established it informs decision-making about service organization, allocation of funds,

and policy priorities in child mental health [20]. For service evaluations and benchmarking to be fair and meaningful, metrics and measurement approaches must be comparable. This study suggests that reliable or meaningful change indicators are not inherently comparable if drawn from different outcome domains or measures. Aggregating change into a single composite indicator risks foregoing the benefits of multidimensional outcome measurement by masking differences in treatment impact on different domains. Making maximum use of all available data and exploring what might drive inconsistencies between these data would enable more nuanced insights into what treatments work; for whom; and with regards to which outcome [93]. Emerging global standards for routine outcome measurement aim to promote harmonization, but require widespread uptake by mental health systems and services to be successful. Such initiatives can help focus resources and interest on a set of recommended measures, which can then be calibrated thoroughly against one another, and for which reliability, validity, and sensitivity to change can be studied in detail [101]. With practice-based evidence increasingly driving decisions about care, the stakes are high and ambiguity in outcomes reporting must be avoided.

Summary

Anxiety and depression are prevalent mental health problems in adolescence. Routinely collected outcome data is key to providing an effective and evidence-based clinical response. Those wishing to compare outcomes achieved in different services or systems are typically faced with data obtained through inconsistent measurement approaches. Mitigation strategies include the comparison of symptom-focused metrics, such as rates of remission or recovery, or the aggregation of results obtained from different measures or outcome domains into a comparable core metric, such as the reliable change index. The implications of either approach for judging treatment success are not well understood.

This study compared meaningful change between two self-report measures of internalizing symptoms; two brief self-report measures of psychosocial functioning; and between the outcome domains of symptoms, functioning and progress towards self-defined goals. Meaningful change was defined based on the reliable change index for standardized scales, and based on a meaningful change index for the idiographic goal-based outcome measure. The data originated from a naturalistic sample of 1,641 adolescents with moderate or severe anxiety and/or depression symptoms who were treated across 60 child and adolescent mental health services in England. Improvement was observed consistently across all three outcome domains in only 15.6% of cases. A comparison of reliable improvement ratings between domains revealed that 24% of young people showed reliable improved in symptoms, but no improvement in functioning. Inversely, 34.8% of young people showed meaningful progress towards self-defined goals, but no improvement in symptoms.

Focusing outcome reporting exclusively on symptom change risks over- or under-estimating actual treatment effectiveness, while aggregating data from several outcome domains into a single metric can mask informative differences in the number and type of outcomes showing improvement. Instead, mental health systems and services should consider assessing multiple outcomes to gain a more balanced picture of the changes achieved, and should draw on emerging standards for the selection of outcome domains, measurement instruments, and assessment time points, to ensure that their data can add to a growing, harmonized evidence base that allows for meaningful comparisons.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10578-021-01149-y.

Acknowledgements We are grateful to all services that participated in the Children and Young People's Improving Access to Psychological Therapies (CYP IAPT) initiative for providing the data underpinning this study. We are grateful to the Child Outcomes Research Consortium (CORC) central team researchers who helped with preparing the data for analysis: Dr. Luís Costa da Silva, Rory Lawless, Meera Patel, Anisatu Rashid, and Benjamin Ritchie. We would also like to thank Prof. Célia Sales for her valuable suggestions on an earlier version of this manuscript.

Funding This study was funded through an IMPACT Studentship awarded to Karolin Krause by University College London and the Anna Freud National Centre for Children and Families. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declarations

Conflict of Interest KRK was involved in the development of the ICHOM standard set of outcomes for child and youth anxiety and depression and received personal fees from ICHOM during the development of the set (October 2018-March 2020). The ICHOM standard set is available free of charge and there is no financial conflict of interest. KRK is involved in the development of a core outcome set for adolescent depression clinical trials as part of the International Network for Research Outcomes in Adolescent Depression Studies (IN-ROADS) initiative. JEC and MW were involved in the programme of service transformation that generated the data that this manuscript draws on. MW led the outcomes and evaluation group that agreed the approach to measurement used in the initiative. MW is now Head of the new Mental Health Priority Area at the Wellcome Trust, which may be developing core outcome sets in mental health in the future. MW has been involved in the development of the Current View Tool, which is a freely available measure and there is no financial conflict of interest. MW was previously Head of the Child Outcomes Research Consortium (CORC), which advises on measurement in child mental health, and she was an advisor to NHS England on informatics. RS has no conflicts to disclose.

References

1. Bor W, Dean AJ, Najman J, Hayatbakhsh R (2014) Are child and adolescent mental health problems increasing in the 21st century? A systematic review. Aust N Z J Psychiatry 48:606–616. https://doi.org/10.1177/0004867414533834
2. Polanczyk GV, Salum GA, Sugaya LS et al (2015) Annual research review: a meta-analysis of the worldwide prevalence of mental disorders in children and adolescents. J Child Psychol Psychiatry 56:345–365. https://doi.org/10.1111/jcpp.12381
3. World Health Organization (2019) Adolescent mental health. https://www.who.int/news-room/fact-sheets/detail/adolescent-mental-health. Accessed 3 Dec 2019
4. Clayborne ZM, Varin M, Colman I (2019) Systematic review and meta-analysis: adolescent depression and long-term psychosocial outcomes. J Am Acad Child Adolesc Psychiatry 58:72–79. https://doi.org/10.1016/j.jaac.2018.07.896
5. Bruce SE, Yonkers KA, Otto MW et al (2005) Influence of psychiatric comorbidity on recovery and recurrence in generalized anxiety disorder, social phobia, and panic disorder: a 12-year prospective study. Am J Psychiatry 162:1179–1187. https://doi.org/10.1176/appi.ajp.162.6.1179
6. Holsen I, Birkeland MS (2017) Course, predictors, and outcomes of depressed mood in a cohort of norwegian adolescents followed from age 13 into adulthood age 30. Emerg Adulthood 5:3–15. https://doi.org/10.1177/2167696816635810
7. Jonsson U, Bohman H, von Knorring L et al (2011) Mental health outcome of long-term and episodic adolescent depression: 15-year follow-up of a community sample. J Affect Disord 130:395–404. https://doi.org/10.1016/j.jad.2010.10.046
8. Weisz JR, Kuppens S, Ng MY et al (2017) What five decades of research tells us about the effects of youth psychological therapy: a multilevel meta-analysis and implications for science and practice. Am Psychol 72:79–117
9. Eckshtain D, Kuppens S, Ugueto A et al (2020) Meta-analysis: 13-year follow-up of psychotherapy effects on youth depression. J Am Acad Child Adolesc Psychiatry 59:45–63. https://doi.org/10.1016/j.jaac.2019.04.002
10. Oud M, de Winter L, Vermeulen-Smit E et al (2019) Effectiveness of CBT for children and adolescents with depression: a systematic review and meta-regression analysis. Eur Psychiatry 57:33–45. https://doi.org/10.1016/j.eurpsy.2018.12.008
11. Weisz JR, Kuppens S, Ng MY et al (2019) Are psychotherapies for young people growing stronger? Tracking trends over time for youth anxiety, depression, attention-deficit/hyperactivity disorder, and conduct problems. Perspect Psychol Sci 14:216–237. https://doi.org/10.1177/1745691618805436
12. Zhou X, Hetrick SE, Cuijpers P et al (2015) Comparative efficacy and acceptability of psychotherapies for depression in children and adolescents: a systematic review and network meta-analysis. World Psychiatry 14:207–222. https://doi.org/10.1002/wps.20217
13. Chorpita BF, Daleiden EL (2009) Mapping evidence-based treatments for children and adolescents: application of the distillation and matching model to 615 treatments from 322 randomized trials. J Consult Clin Psychol 77:566–579. https://doi.org/10.1037/a0014565
14. Silverman WK, Hinshaw SP (2008) The second special issue on evidence-based psychosocial treatments for children and adolescents: a 10-year update. J Clin Child Adolesc Psychol 37:1–7. https://doi.org/10.1080/15374410701817725
15. Weisz JR, Krumholz LS, Santucci L et al (2015) Shrinking the gap between research and practice: tailoring and testing youth psychotherapies in clinical care contexts. Annu Rev Clin Psychol 11:139–163. https://doi.org/10.1146/annurev-clinpsy-032814-112820
16. Bear HA, Edbrooke-Childs J, Norton S et al (2020) Systematic review and meta-analysis: outcomes of routine specialist mental health care for young people with depression and/or anxiety. J Am Acad Child Adolesc Psychiatry 59:810–841. https://doi.org/10.1016/j.jaac.2019.12.002
17. Metz MJ, Veerbeek MA, Twisk JWR et al (2019) Shared decision-making in mental health care using routine outcome monitoring: results of a cluster randomised-controlled trial. Soc Psychiatry Psychiatr Epidemiol 54:209–219. https://doi.org/10.1007/s00127-018-1589-8
18. Boswell JF, Kraus DR, Miller SD, Lambert MJ (2015) Implementing routine outcome monitoring in clinical practice: benefits, challenges, and solutions. Psychother Res 25:6–19. https://doi.org/10.1080/10503307.2013.817696
19. Delgadillo J, Overend K, Lucock M et al (2017) Improving the efficiency of psychological treatment using outcome feedback technology. Behav Res Ther 99:89–97. https://doi.org/10.1016/j.brat.2017.09.011
20. Garralda EM (2009) Accountability of specialist child and adolescent mental health services. Br J Psychiatry 194:389–391. https://doi.org/10.1192/bjp.bp.108.059477
21. Krause KR, Chung S, Adewuya AO et al (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
22. Sharples E, Qin C, Goveas V et al (2017) A qualitative exploration of attitudes towards the use of outcome measures in child and adolescent mental health services. Clin Child Psychol Psychiatry 22:219–228. https://doi.org/10.1177/1359104516652929
23. Stasiak K, Parkin A, Seymour F et al (2013) Measuring outcome in child and adolescent mental health services: consumers' views of measures. Clin Child Psychol Psychiatry 18:519–535. https://doi.org/10.1177/1359104512460860
24. Norman S, Dean S, Hansford L, Ford T (2014) Clinical practitioner's attitudes towards the use of routine outcome monitoring within child and adolescent mental health services: a qualitative study of two child and adolescent mental health services. Clin Child Psychol Psychiatry 19:576–595. https://doi.org/10.1177/1359104513492348
25. Spitzer RL, Kroenke K, Williams JBW (1999) Validation and utility of a self-report version of PRIME-MD: The PHQ Primary Care Study. J Am Med Assoc 282:1737–1744. https://doi.org/10.1001/jama.282.18.1737
26. Krause KR, Bear HA, Edbrooke-Childs J, Wolpert M (2019) Review: what outcomes count? A review of outcomes measured for adolescent depression between 2007 and 2017. J Am Acad Child Adolesc Psychiatry 58:61–71. https://doi.org/10.1016/j.jaac.2018.07.893

27. Childs J, Deighton J, Wolpert M (2013) Defining and measuring mental health and wellbeing: a response mode report requested by the Department of Health for the Policy Research Unit in the Health of Children. Young People and Families, London
28. Kazdin AE (1999) The meanings and measurement of clinical significance. J Consult Clin Psychol 67:332–339. https://doi.org/10.1037/0022-006X.67.3.332
29. Blanton H, Jaccard J (2006) Arbitrary metrics in psychology. Am Psychol 61:27–41. https://doi.org/10.1037/0003-066X.61.1.27
30. The Lancet Psychiatry (2020) Measuring success: the problem with primary outcomes. Lancet Psychiatry 7:1. https://doi.org/10.1016/S2215-0366(19)30483-3
31. Zimmerman M, McGlinchey JB, Posternak MA et al (2008) Remission in depressed outpatients: more than just symptom resolution? J Psychiatr Res 42:797–801. https://doi.org/10.1016/j.jpsychires.2007.09.004
32. Kazdin AE (2006) Arbitrary metrics: implications for identifying evidence-based treatments. Am Psychol 61:42–49. https://doi.org/10.1037/0003-066X.61.1.42
33. Jacob J (2019) Moving toward a better understanding of idiographic outcome measurement: a commentary on Lloyd, Duncan, and Cooper (2019). Clin Psychol Sci Pract. https://doi.org/10.1111/cpsp.12287
34. Wolpert M, Jacob J, Napoleone E et al (2016) Child- and parent-reported outcomes and experience from child and young people's mental health services 2011–2015. CAMHS Press, London
35. Sales CMD (2017) Seeing the person in the patient: making the case for individualized PROMs in mental health care. Curr Psychiatry Rev 13:184–187. https://doi.org/10.2174/1573400513666170505111736
36. Sales CMD, Alves PCG (2016) Patient-centered assessment in psychotherapy: a review of individualized tools. Clin Psychol Sci Pract 23:265–283. https://doi.org/10.1111/cpsp.12162
37. Wolpert M (2017) Commentary: why measuring clinical change at the individual level is challenging but crucial: commentary on Jensen and Corralejo (2017). Child Adolesc Ment Health 22:167–169. https://doi.org/10.1111/camh.12228
38. Wolpert M, Görzig A, Deighton J et al (2015) Comparison of indices of clinically meaningful change in child and adolescent mental health services: difference scores, reliable change, crossing clinical thresholds and "added value": an exploration using parent rated scores on the SDQ. Child Adolesc Ment Health 20:94–101. https://doi.org/10.1111/camh.12080
39. Becker KD, Chorpita BF, Daleiden EL (2011) Improvement in symptoms versus functioning: how do our best treatments measure up? Adm Policy Ment Heal Ment Heal Serv Res 38:440–458. https://doi.org/10.1007/s10488-010-0332-x
40. Brookman-Frazee L, Haine RA, Garland AF (2006) Innovations: child and adolescent psychiatry: measuring outcomes of real-world youth psychotherapy: whom to ask and what to ask? Psychiatr Serv 57:1373–1375. https://doi.org/10.1176/ps.2006.57.10.1373
41. McKnight PE, Kashdan TB (2009) The importance of functional impairment to mental health outcomes: a case for reassessing our goals in depression treatment research. Clin Psychol Rev 29:243–259. https://doi.org/10.1016/j.cpr.2009.01.005
42. Karpenko V, Owens JS (2013) Adolescent psychotherapy outcomes in community mental health: how do symptoms align with target complaints and perceived change? Community Ment Health J 49:540–552. https://doi.org/10.1007/s10597-012-9515-0
43. Edbrooke-Childs J, Jacob J, Law D et al (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
44. Costello EJ, Shugart MA (1992) Above and below the threshold: severity of psychiatric symptoms and functional impairment in a pediatric sample. Pediatrics 90:359–368
45. Simonoff E, Pickles A, Meyer JM et al (1997) The Virginia twin study of adolescent behavioral development: influences of age, sex, and impairment on rates of disorder. Arch Gen Psychiatry 54:801–808. https://doi.org/10.1001/archpsyc.1997.01830210039004
46. Jensen SA, Corralejo SM (2017) Measurement issues: large effect sizes do not mean most people get better: clinical significance and the importance of individual results. Child Adolesc Ment Health 22:163–166. https://doi.org/10.1111/camh.12203
47. Costa da Silva L, Wolpert M (2018) Outcome measures in child and youth mental health services: results from NHS England survey. Child Outcomes Research Consortium (CORC), London
48. Deighton J, Tymms P, Vostanis P et al (2013) The development of a school-based measure of child mental health. J Psychoeduc Assess 31:247–257. https://doi.org/10.1177/0734282912465570
49. Wolpert M, Fugard AJB, Deighton J, Görzig A (2012) Routine outcomes monitoring as part of children and young people's improving access to psychological therapies (CYP IAPT): improving care or unhelpful burden? Child Adolesc Ment Health 17:129–130. https://doi.org/10.1111/j.1475-3588.2012.00676.x
50. Jones M, Hopkins K, Kyrke-Smith R et al (2013) Current view tool: completion guide. CAMHS Press, London
51. National Health Service (2018) Governance arrangements for research ethics committees: 2018 edition. https://www.hra.nhs.uk/planning-and-improving-research/policies-standards-legislation/governance-arrangement-research-ethics-committees/. Accessed 11 Jun 2019
52. World Health Organization (2018) International classification of diseases for mortality and morbidity statistics (11th revision). https://icd.who.int/browse11/l-m/en. Accessed 15 Oct 2019
53. Chorpita BF, Yim L, Moffitt C et al (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
54. American Psychiatric Association (1994) Diagnostic and statistical manual of mental disorders, 4th edn. American Psychiatric Association, Washington, DC
55. Chorpita BF, Moffitt CE, Gray J (2005) Psychometric properties of the revised child anxiety and depression scale in a clinical sample. Behav Res Ther 43:309–322. https://doi.org/10.1016/j.brat.2004.02.004
56. de Ross RL, Gullone E, Chorpita BF (2002) The revised child anxiety and depression scale: a psychometric investigation with australian youth. Behav Chang 19:90–101. https://doi.org/10.1375/bech.19.2.90
57. Kösters MP, Chinapaw MJM, Zwaanswijk M et al (2015) Structure, reliability, and validity of the revised child anxiety and depression scale (RCADS) in a multi-ethnic urban sample of Dutch children. BMC Psychiatry 15:132. https://doi.org/10.1186/s12888-015-0509-7
58. Esbjørn BH, Sømhovd MJ, Turnstedt C, Reinholdt-Dunne ML (2012) Assessing the revised child anxiety and depression scale (RCADS) in a national sample of Danish youth aged 8–16 years. PLoS ONE 7:1–5. https://doi.org/10.1371/journal.pone.0037339
59. Goodman R, Meltzer H, Bailey V (1998) The strengths and difficulties questionnaire: a pilot study on the validity of the self-report version. Eur Child Adolesc Psychiatry 7:125–130
60. Goodman R (1997) The strengths and difficulties questionnaire: a research note. J Child Psychol Psychiat 38:581–586
61. Achenbach TM, Becker A, Döpfner M et al (2008) Multicultural assessment of child and adolescent psychopathology with ASEBA and SDQ instruments: research findings, applications,

and future directions. J Child Psychol Psychiatry Allied Discip 49:251–275. https://doi.org/10.1111/j.1469-7610.2007.01867.x
62. Goodman R (2001) Psychometric properties of the strengths and difficulties questionnaire (SDQ). J Am Acad Child Adolesc Psychiatry 40:1337–1345. https://doi.org/10.1097/00004583-200111000-00015
63. Muris P, Meesters C, van den Berg F (2003) The strengths and difficulties questionnaire (SDQ): further evidence for its reliability and validity in a community sample of Dutch children and adolescents. Eur Child Adolesc Psychiatry Adolesc psychiatry 12:1–8
64. Goodman A, Goodman R (2009) Strengths and difficulties questionnaire as a dimensional measure of child mental health. J Am Acad Child Adolesc Psychiatry 48:400–403. https://doi.org/10.1097/CHI.0b013e3181985068
65. Goodman R (1999) The extended version of the strengths and difficulties questionnaire as a guide to child psychiatric caseness and consequent burden. J Child Psychol Psychiatry 40:791–799
66. Miller S, Duncan B, Brown J et al (2003) The outcome rating scale: a preliminary study of the reliability, validity, and feasibility of a brief visual analog measure. J Br Ther 2:91–100
67. Duncan BL, Miller SD, Sparks JA et al (2003) The session rating scale: preliminary psychometric properties of a "working" alliance measure. J Br Ther 3:3–12
68. Bringhurst MDL, Watson CW, Miller SD, Duncan BL (2006) The reliability and validity of the outcome rating scale: a replication study of a brief clinical measure. J Br Ther 5:23–30
69. Campbell AG, Hemsley S (2009) The validity of the outcome rating scale and session rating scale in psychological practice. Clin Psychol 13:1–10
70. Casey P, Patalay P, Deighton J et al (2019) The child outcome rating scale: validating a four-item measure of psychosocial functioning in community and clinic samples of children aged 10–15. Eur Child Adolesc Psychiatry. https://doi.org/10.1007/s00787-019-01423-4
71. Law D (2006) Goal Based Outcomes (GBO): some useful information. CAMHS Press, London
72. Wolpert M, Cheng H, Deighton J (2015) Measurement issues: review of four patient reported outcome measures: SDQ, RCADS, C/ORS and GBO: their strengths and limitations for clinical use and service evaluation. Child Adolesc Ment Health 20:63–70. https://doi.org/10.1111/camh.12065
73. Law D, Jacob J (2015) Goals and goal based outcomes (GBOs): some useful information, 3rd edn. CAMHS Press, London
74. Law D (2018) Goals and goal-based outcomes (GBOs): goal progress chart. Version 2.0. CAMHS Press, London
75. Jacob J, Edbrooke-Childs J, Lloyd C et al (2018) Measuring outcomes using goals. In: Cooper M, Law D (eds) Working with goals in psychotherapy and counselling. Oxford University Press, Oxford, pp 111–138
76. Jacobson NS, Truax P (1991) Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. J Consult Clin Psychol 59:12–19. https://doi.org/10.1037/0022-006X.59.1.12
77. Jacobson NS, Roberts LJ, Berns SB, McGlinchey JB (1999) Methods for defining and determining the clinical significance of treatment effects. J Consult Clin Psychol 67:300–307
78. Office for National Statistics (2006) Review of the dissemination of health statistics: confidentiality guidance. Office for National Statistics, London
79. McNemar Q (1947) Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 12:153–157
80. Cohen J (1960) A Coefficient Of Agreement For Nominal scales. Educ Psychol Meas 20:37–46. https://doi.org/10.1177/001316446002000104
81. Fleiss JL (1981) Statistical methods for rates and proportions. Wiley, New York
82. Lee W, Jones L, Goodman R, Heyman I (2005) Broad outcome measures may underestimate effectiveness: an instrument comparison study. Child Adolesc Ment Health 10:143–144. https://doi.org/10.1111/j.1475-3588.2005.00350.x
83. Wolpert M, Rutter H (2018) Using flawed, uncertain, proximate and sparse (FUPS) data in the context of complexity: learning from the case of child mental health. BMC Med 16:1–11. https://doi.org/10.1186/s12916-018-1079-6
84. Patrick DL, Deyo RA (1989) Generic and disease-specific measures in assessing health status and quality of life. Med Care 27:S217–S232
85. Fried EI, Nesse RM (2014) The impact of individual depressive symptoms on impairment of psychosocial functioning. PLoS ONE. https://doi.org/10.1371/journal.pone.0090311
86. Canino G, Fisher PW, Alegria M, Bird HR (2013) Assessing child impairment in functioning in different contexts: implications for use of services and the classification of psychiatric disorders. Open J Med Psychol 2:29–34. https://doi.org/10.4236/ojmp.2013.21006
87. Hurn J, Kneebone I, Cropley M (2006) Goal setting as an outcome measure: a systematic review. Clin Rehabil 20:756–772. https://doi.org/10.1177/0269215506070793
88. Ashworth M, Robinson SI, Godfrey E et al (2005) Measuring mental health outcomes in primary care: the psychometric properties of a new patient-generated outcome measure, "PSYCHLOPS" ('psychological outcome profiles'). Prim Care Ment Heal 3:261–270
89. Elliott R, Wagner J, Sales CMD et al (2016) Psychometrics of the personal questionnaire: a client-generated outcome measure. Psychol Assess 28:263–278. https://doi.org/10.1037/pas0000174
90. Wolpert M, Ford T, Trustam E et al (2012) Patient-reported outcomes in child and adolescent mental health services (CAMHS): use of idiographic and standardized measures. J Ment Heal 21:165–173. https://doi.org/10.3109/09638237.2012.664304
91. Jacob J, Edbrooke-Childs J, Law D, Wolpert M (2017) Measuring what matters to patients: using goal content to inform measure choice and development. Clin Child Psychol Psychiatry 22:170–186. https://doi.org/10.1177/1359104515615642
92. O'Keeffe S, Martin P, Target M, Midgley N (2019) "I just stopped going": a mixed methods investigation into types of therapy dropout in adolescents with depression. Front Psychol 10:1–14. https://doi.org/10.3389/fpsyg.2019.00075
93. De Los RA, Kazdin AE (2008) When the evidence says, "yes, no, and maybe so." Curr Dir Psychol Sci 17:47–51. https://doi.org/10.1111/j.1467-8721.2008.00546.x
94. Duong MT, Cruz RA, King KM et al (2016) Twelve-month outcomes of a randomized trial of the positive thoughts and action program for depression among early adolescents. Prev Sci 17:295–305. https://doi.org/10.1007/s11121-015-0615-2
95. Cairns AJ, Kavanagh DJ, Dark F, McPhail SM (2019) Goal setting improves retention in youth mental health: a cross-sectional analysis. Child Adolesc Psychiatry Ment Health 13:1–8. https://doi.org/10.1186/s13034-019-0288-x
96. Revicki D, Hays RD, Cella D, Sloan J (2008) Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. J Clin Epidemiol 61:102–109. https://doi.org/10.1016/j.jclinepi.2007.03.012
97. Reeve BB, Wyrwich KW, Wu AW et al (2013) ISOQOL recommends minimum standards for patient-reported outcome measures used in patient-centered outcomes and comparative effectiveness research. Qual Life Res 22:1889–1905. https://doi.org/10.1007/s11136-012-0344-y

98. McGlothlin AE, Lewis RJ (2014) Minimal clinically important difference: defining what really matters to patients. JAMA 312:1342–1343. https://doi.org/10.1001/jama.2014.13128
99. Canino G, Costello EJ, Angold A (1999) Assessing functional impairment and social adaptation for child mental health services research: a review of measures. Ment Health Serv Res 1:93–108. https://doi.org/10.1023/A:1022334303731
100. Rockwood K, Joyce B, Stolee P (1997) Use of goal attainment scaling in measuring clinically important change in cognitive rehabilitation patients. J Clin Epidemiol 50:581–588. https://doi.org/10.1016/S0895-4356(97)00014-0
101. Sechrest L, McKnight PE, McKnight K (1996) Calibration of measures for psychotherapy outcome studies. Am Psychol 51:1065–1071. https://doi.org/10.1007/978-1-4614-6435-8_452-3