Received: 27 November 2019 | Accepted: 25 February 2022

DOI: 10.1002/jclp.23344

RESEARCH ARTICLE

WILEY

The goals form: Reliability, validity, and clinical utility of an idiographic goal-focused measure for routine outcome monitoring in psychotherapy

Mick Cooper[1] | Dan Xu[1,2]

[1]School of Psychology, University of Roehampton, London, UK

[2]Department of Psychology, Zhejiang University of Technology, Hangzhou, Zhejiang, China

Correspondence
Mick Cooper, School of Psychology, University of Roehampton, London SW15 4JD, UK.
Email: mick.cooper@roehampton.ac.uk

Abstract

Objective: This study aimed to assess the reliability, validity, and clinical utility of an idiographic, goal-focused patient-reported outcome measure: The Goals Form.

Methods: Data were analyzed from 88 participants, across three samples, who had participated in collaborative–integrative psychotherapy at university-based clinics in the UK. The samples were approximately 70% female with mean age of 30 years old.

Results: The psychometric properties of the Goals Form were generally good. Noncompletion of individual items was low, temporal stability tended to be at target levels, and mean change scores showed moderate to good convergent validity against measures of psychological distress. The measure appeared sensitive to change in psychotherapy and was experienced by most patients as helpful.

Conclusions: The Goals Form shows acceptable psychometric and clinical properties for routine outcome monitoring in psychotherapy.

KEYWORDS

goals, idiographic, outcome and process assessment, patient-generated measures, routine outcome monitoring

1 | INTRODUCTION

Routine outcome monitoring (ROM) is a major development in the psychotherapy field (Lambert et al., 2018), and has the potential to serve three key functions. First, it may improve individual clinical outcomes—particularly for "not on track" patients (Bickman et al., 2011; Lambert et al., 2018)—though these claims have been contested (Bergman et al., 2018; Kendrick et al., 2016). Second, it may enhance patients' experiences of the therapeutic process: helping them, for instance, feel more aware of their problems or goals or giving them feedback on progress (Di Malta et al., 2019). Third, it can evidence outcomes at the service and treatment level, informing decisions about commissioning and policy-making.

To date, ROM systems have almost exclusively relied upon nomothetic patient-reported outcome measures (PROMs). Nomothetic PROMs are typically brief, with items that are predefined and predetermined: consistent across patients (Sales & Alves, 2016). Such qualities make nomothetic PROMs particularly suited to service and treatment level evaluation, where outcomes need to be compared (for instance, across sites, or against benchmarks). However, nomothetic PROMs may not capture the specific changes that are of greatest importance to individual patients, or differences in how particular items are interpreted (Krause et al., 2019; Moltu et al., 2017; Sales et al., 2022).

An alternative approach to ROM, therefore, is the use of *idiographic* PROMs (I-PROMs). Here, patients construct—and rate progress against—their own items, within a standardized questionnaire format (Di Malta et al., 2019; Sales et al., 2022). Such individualization allows patients to define their own psychotherapy foci, enabling the broadest possible array of value systems and conceptualizations of treatment success.

I-PROMs can be either *problem-focused* or *goal-focused* (Lloyd et al., 2019; Sales et al., 2022). Problem-focused I-PROMs invite patients to identify the difficulties, issues, or concerns that they would like to address. Goal-focused I-PROMs invite them to identify the objectives or "end-states" that they would like to attain.

Problem-focused I-PROMs, such as the PSYCHLOPS and the Personal Questionnaire, have demonstrated excellent psychometric properties, and have been successfully applied in the psychotherapy domain (e.g., Ashworth et al., 2005; Elliott et al., 2016; Sales et al., 2021). However, as Lloyd et al. (2019) argue, goal-oriented I-PROMs are supported by several additional lines of psychological and psychotherapeutic research. First, there is a large body of psychological data to show that goal setting and goal monitoring procedures can enhance task performance (Locke & Latham, 2002; Locke et al., 1981), with effect sizes (*d*) of 0.34 (Epton et al., 2017), and 0.40 (Harkin et al., 2016), respectively, across a range of behavioral outcomes. Research suggests that these effects may come about because goal-oriented processes direct individuals' attention to identified goals, mobilize effort, support persistence, and motivate individuals to develop strategies for goal attainment (Locke & Latham, 2002; Locke et al., 1981). Second, it is a well-established finding that patients' outcomes in psychotherapy are improved when patients and psychotherapists agree on the goals for the work (Tryon et al., 2018)—a consensus that may be enhanced through explicit goal-oriented practices. Third, goal setting is desired by a majority of patients (Cooper et al., 2019). Given that preference accommodation is associated with improved outcomes and retention (Swift et al., 2018), it is likely that goal-oriented practices, overall, will have a positive effect. Fourth, in contrast to problem-focused I-PROMs, goal-focused I-PROMs allow patients to set "approach" objectives (i.e., what they want), as well as "avoidance" objectives (i.e., what they do not want). Psychological and psychotherapy research suggests that the former may be more effective regulatory devices, and are associated with improved psychotherapy outcomes (Elliot & Church, 2002).

Lloyd et al. (2019) conducted a systematic review of goal-focused I-PROMs that had been used in psychotherapy. They found nine measures that fell into three discrete categories. First were multidimensional tools—such as Personal Project Analysis (PPA, Little, 1983) and the Motivational Structure Questionnaire (Cox & Klinger, 2021)—that invite patients to establish goals through relatively structured processes, and then to rate them on a range of dimensions. Second was Goal Assessment Scaling (GAS, Kiresuk & Sherman, 1968), which invites patients to set out, and then rate, expected levels of outcomes on a set of identified goals. Third was two simple rating

forms, the Goals Form (Cooper, 2015) and the Goal-Based Outcome (GBO) tool (Law, 2019), which invite patients to set goals through a relatively brief and unstructured goal-setting process, and then to rate progress for each goal on a single dimension. As a consequence of the brevity and simplicity of this third category of measures, Lloyd et al. suggested that they were the form of goal-focused I-PROM most suited to ROM.

The focus of the present article is on the psychometric properties of the Goals Form: the GBO Tool being primarily designed for use with children and adolescents (Jacob et al., 2021; Law, 2019). Although preliminary evidence for the reliability and validity of the Goals Form is cited by Lloyd et al. (2019)—and Michael (2014) analyzed Goals Form data for her doctoral thesis—the Goal Form's psychometric properties have yet to be analyzed, or reported on, at a level of statistical rigor or depth. Evidence of the clinical utility of the Goals Form is also limited; although a recent qualitative study of goal-oriented practices—based around the use of the Goals Form—found that these methods could be experienced by patients as motivating and mobilizing, helping them to focus efforts and establish more realistic treatment expectations (Di Malta et al., 2019). Concurrently, however, Di Malta et al. found "clear evidence" that some patients could experience goal-oriented practices as having unwanted or negative effects: in particular, an uncertainty around what goals to set. Overall, in Di Malta et al.'s study, 15 of 22 patients (68.2%) indicated that they found the Goals Form helpful.

In assessing the psychometric properties of a measure, it is essential to specify the measurement model on which it is based (Bollen & Diamantopoulos, 2017). This is so that the correct tests can be selected for their validation. The two principal types of measurement models are *reflective* and *formative* (Bollen & Bauldry, 2011; Diamantopoulos & Winklhofer, 2001; Edwards & Bagozzi, 2000; Jarvis et al., 2003). In the reflective model, items are seen as reflecting—and being caused by—a hypothesized latent construct. This is the measurement model underlying classical test theory (e.g., Nunnally & Bernstein, 1994) and all nomothetic PROMs. For instance, on the Generalized Anxiety Disorder-7 (GAD-7) (Spitzer et al., 2006), the patient's response to the item "[I have been bothered by] Feeling nervous, anxious or on edge?" would be considered reflective of, and caused by, the existence of an underlying problem with generalized anxiety. By contrast, in the *formative* measurement model, the observed indicators form—in the sense that they cause—the hypothesized construct. For instance, exposure to stress is determined by a set of stressful life events, such as getting married or divorced, losing a loved one, changing a job. Here, there is no latent "exposure to stress" construct that is driving the observed indicators. Rather, the causal indicators, together, create this construct.

Where measurement is formative, certain key elements of measure validation within classical test theory—in particular, the necessity to demonstrate internal consistency, along with unidimensionality—are obviated (Diamantopoulos et al., 2008). This is because, in contrast to reflective measures, we are not assuming, or requiring, that our indicators co-vary. With a reflective measure such as the GAD-7, for instance, if patients' scores on "Feeling nervous, anxious or on edge" did not co-vary with their scores on the item "Trouble relaxing?" this would suggest that the measure is not adequately tapping an underlying variable. But, in the example of exposure to stress, we would not necessarily expect such factors as getting married and losing a loved one to co-vary; and, if they did not, this would not make our overall indicator of exposure to stress any less valid. In this respect, then, combined or averaged scores also have a different meaning in formative measures, as compared with a reflective model. In the latter, they are the closest, most error-free estimation we can make of the latent variable. But, in a formative model, they are a summary of the effects of several variables (Bollen, 2011), and have no independent standing outside of this.

As with Sales et al. (2022), we view items on the Goals Form I-PROM as having the potential to be reflective indicators, formative indicators, or a combination of the two. That is, Goals Form items, like items on a nomothetic PROM, may be reflective of a latent variable. For instance, an individual who has an underlying desire to improve their self-esteem may subsequently establish such psychotherapy goals as, "I want to feel better about who I am" and "I want to be more confident." Improvements in underlying self-esteem as a goal may then be reflected in improvements on the ratings of these two goal items. On the other hand, however, an individual may set Goals Form items such as "I want to improve my relationship with my mother" and "I want to stop smoking." Here, ratings

on each item can be combined to indicate an overall sense of achievement, but they are not driven by some single, underlying variable. There is, of course, another possibility: an individual may set the four-goal items above, simultaneously. This would give a combination of both reflective and formative indicators. However, the combined score would still create a unified, latent variable: the extent of an individual's aggregate achievement on key personal goals.

Given this specification—that items on the Goals Form may be formative as well as reflective—we believe, as with Sales et al. (2022), that only tests appropriate to both models should be used for validating I-PROMs. While our analysis of the Goals Form, therefore, is based on Elliott et al.'s (2016) framework for examining the psychometric properties of an I-PROM (the Personal Questionnaire), we do not present data on internal consistency and dimensionality (Edwards & Bagozzi, 2000). In addition, while we present and analyze mean goal item scores at the session-by-session level, these are considered as "aggregate session scores," which could either be reflective or formative of the unified, latent variable of overall goal-achievement.

The principal qualities that we investigate are as follows: (a) Normative: That typical quantitative characteristics of Goals Form scores can be established, including a number of goal items, length of time goals are "active," and goal scores at baseline and endpoint. (b) Acceptability: That there will be high rates of completed items (<10%, based on acceptable levels of prorating, CORE System Trust, 2015). (c) Temporal structure: Goals Form scores will be consistent over time ($r > 0.70$), showing large test–retest correlations from assessment to the first session and from the last session to follow-up review, and high levels of statistical nonindependence in the form of session-to-session autocorrelations. (d) Construct validity: Goals Form scores will show moderate-to-strong negative correlations (in the −0.40 to −0.60 range) with nomothetic outcome measures of psychological distress, but not so strong as to be redundant (<−0.70); in addition, we predicted nonsignificant correlations between Goals Form scores at baseline with a measure of discriminant validity (ratings of session satisfaction). (e) Sensitivity to change: Goals Form scores will indicate change over the course of psychotherapy, showing large pre–post effects and statistically reliable change (consistent with established nomothetic PROMs); in addition, there will be classification consistency of reliable change between the Goals Form and nomothetic PROMs (Cohen's $\kappa > 0.40$). It is also hypothesized that the Goals Form will be sensitive enough to capture change at the session-by-session level, with Goals Form score increasing step-by-step over the course of psychotherapy. Two methods were utilized to test this hypothesis. First, Goals Form score at session T were hypothesized to be significantly larger than scores at session T-1. Second, we hypothesized that Goals Form scores would increase with the number of sessions: that is, a positive, within-patient correlation would be found between session number and Goals Form score. (f) Clinical utility: On average, patients would rate the Goals Form as helpful to their psychotherapeutic work.

2 | METHODS

2.1 | Participants

Data for this study came from three independently collected samples at free, university-located psychotherapy research clinics in the UK. The combined number of patients was 88, with 1736 observations in total. Although this is a relatively small sample for a measure development study, there is no consensually agreed minimum for such research (Anthoine et al., 2014); and the present subject to item ratio is greater than 10 for the maximum number of Goals Form items ($n = 8$), and greater than 20 for the mean number of Goals Form items ($n = 4$). This exceeds the criteria recommended by Nunnally (1994) and is greater than the average subject to item ratio for other patient-reported outcome measure development studies (Anthoine et al., 2014). In addition, throughout our analyses, we conduct within-participant, repeated measures tests. This gives over 300 observations per test, which would be considered a "good" sample in cross-sectional design (Anthoine et al., 2014).

As procedures varied at each of the sites—including intake procedures, length of therapy, and measures used—we analyzed each of the datasets independently. This also allowed us to triangulate findings from across sites.

The first sample consisted of 15 adults (>18 years old) who had attended a clinic in Scotland between March 2010 and January 2014, the "Scottish sample" (Table 1). Ten of these participants were female, four

TABLE 1 Overview of study sample characteristics

	Scottish (n = 15)	Multisite (n = 25)	London (n = 48)
Years data collected	03/2010–01/2014	01/2013–12/2013	01/2016–06/2019
Age (years), mean ± *SD*	32.9 ± 10.7	30.8 ± 12.3	29.6 ± 12.0
Sessions (mean ± *SD*)	22.2 ± 19.8	15.4 ± 8.2	18.4 ± 6.3
Gender, *n* (%)			
Female	10 (66.7%)	18 (72.0%)	33 (68.8%)
Male	4 (26.7%)	7 (28.0%)	12 (25.0%)
Ethnicity			
BME	NA	4 (16.0%)	15 (31.3%)
White European	NA	21 (84.0%)	29 (60.4%)
Disability			
Disabled	NA	3 (12.0%)	6 (12.5%)
Nondisabled	NA	21 (84.0%)	37 (77.1%)
Student			
In full-time education	NA	10 (40.0%)	19 (39.6%)
Not in full-time education	NA	15 (60.0%)	25 (52.1%)
Psychological distress[a]			
Clinical	8 (53.3%)	NA	NA
Nonclinical	7 (46.7%)	NA	NA
Depression[b]			
Moderate depression	NA	9 (36.0%)	11 (22.9%)
Moderately severe depression	NA	9 (36.0%)	13 (27.1%)
Severe depression	NA	7 (28.0%)	23 (47.9%)
Anxiety[c]			
Minimal anxiety	NA	0 (0%)	0 (0%)
Mild anxiety	NA	5 (20.0%)	7 (14.6%)
Moderate anxiety	NA	12 (48.0%)	13 (27.1%)
Severe anxiety	NA	8 (32.0%)	27 (56.3%)

Abbreviations: CORE, Clinical Outcomes in Routine Evaluation Measures; GAD-7, Generalized Anxiety Disorder Scale; PHQ-9, Patient Health Questionnaire Depression Scale; *SD*, standard deviation.

[a]Based on CORE-10 scoring: 0–10 = Nonclinical, 11–40 = Clinical.

[b]Based on PHQ-9 scoring: 10–14 = Moderate depression, 15–19 = Moderately severe depression, 20–27 = Severe depression.

[c]Based on GAD-7 scoring: 0–4 = Minimal anxiety, 5–9 = Mild anxiety, 10–14 = Moderate anxiety, 15–21 = Severe anxiety.

were male, and one unreported, with an average age of 32.9 years (standard deviation [*SD*] = 10.7). At baseline assessment, 53.3% of patients met the criteria for clinical levels of psychological distress on the CORE-OM.

The second, "multisite sample," consisted of 25 adults who had attended one of four UK-based clinics between January and December 2013, as part of a trial of *pluralistic therapy for depression* (Cooper et al., 2015). This was 64.1% of 39 patients who had attended psychotherapy at these clinics, with patients excluded if they had two or less sets of ratings on the Goals Form, or if they had 10 or more goals (indicating that goals were being reset on a regular basis, contrary to Goals Form instructions). Eighteen of the sample were female and 7 were male, 21 were white (84%), 15 were students (60%), with an average age of 30.8 years (*SD* = 12.3). Inclusion criteria were having a Patient Health Questionnaire-9 score consistent with a diagnosis of depression (PHQ-9 ≥ 10, Kroenke et al., 2001). Exclusion criteria were severe mental health conditions, including individuals experiencing psychosis, severe personality disorders, or drug and alcohol addictions.

The third sample consisted of 48 adults who had completed treatment between January 2016 and June 2019 in London, as part of an ongoing evaluation of *Pluralistic Therapy for Depression*: the "London sample." This was 76.2% of the 63 patients who had attended the clinic over this period, with patients excluded if they had not set goals during their psychotherapy. Of this sample, 33 were female (68.8%), 12 were male (25.0%), and 3 unreported (6.4%); 29 were white (60.4%); 19 were students (39.6%); and the average age was 29.6 years (*SD* = 12.0). Eligibility criteria were the same as for the multisite sample.

2.2 | Measures

2.2.1 | Goals Form

The *Goals Form* was developed as a simple, easy to complete I-PROM for ROM in psychotherapy (Cooper, 2015, 201). The form asks patients, in collaboration with their psychotherapist, to identify up to seven goals for therapy (though they can add more)—typically at a first assessment session—and then to rate them on a 1 (*not at all achieved*) to 7 (*completely achieved*) rating scale. The agreed goals are then typed onto a digital copy of the form and printed off (or, in the London sample, presented to the patient on a digital device), such that patients are able to rate the same goals at regular intervals, ideally every session. Over the course of psychotherapy, patients have the opportunity to add, delete, or modify goals (with modified goals, for the purposes of analysis, treated as new goals); and the electronic copy of the Goals Form is revised accordingly. However, patients are discouraged from re-setting goals too frequently—it is expected that most goals will stay in place for at least several sessions. The Goals Form can be downloaded, with instructions for use and scoring, from https://pluralisticpractice.com/tools-and-measures/ (see also Appendix for blank Goals Form). The Goals Form is licensed under the Creative Commons Attribution-NoDerivatives 4.0 International licence so it can be reproduced, used, and distributed without payment of any fee as long as it is not changed and its origin acknowledged (by citing this paper).

For the purposes of ROM, patients and psychotherapists can view changes on individual Goals Form items on a graph or digital device. In addition, an *aggregate session score* can be calculated by averaging all goal scores in the same session for the patient. However, because patients can add or delete goals, changes in the aggregate session score over time may not necessarily be representative of overall psychotherapeutic progress. For instance, if a patient, having completed all their goals in session T, decides to establish a new set of goals in session T + 1, then their aggregate session score in session T + 1 may be lower than in session T, despite their actual achievement of goals. Changes over the course of psychotherapy, however, can be calculated by averaging the differences between baseline and endpoint ratings on each individual goal: the *aggregate change score*. This means that, if a patient does not change their goals, their aggregate change score will be the same as

the difference between their baseline and endpoint aggregate session scores. However, if any goals are added or deleted, these statistics may be different.

2.2.2 | Clinical Outcomes in Routine Evaluation Measures (CORE-OM and CORE-10)

The CORE-OM and CORE-10 measures were used to assess the convergent validity of the Goals Form in the Scottish sample.

The CORE-OM is a measure of psychological distress and comprises 34 items addressing domains of subjective well-being, symptoms, functioning, and risk. Items are scored on a 5-point, 0–4 scale, from *Not at all* to *All or most of the time*. Clinical scores are calculated as the mean of all completed items, multiplied by 10, giving a range from 0 to 40. Internal consistency for the scale in a clinical sample was 0.94 (Barkham et al., 2001). The 1-week test–retest correlation was Spearman's $\rho = 0.90$ ($n = 43$) (Evans et al., 2002).

The CORE-10 is a shortened version of the CORE-OM, with similar scoring procedures. The CORE-10 has good levels of internal reliability, acceptability, and validity (Barkham et al., 2013). The reliable change index of the CORE-10 measure has been calculated as 8 (Barkham et al., 2013).

2.2.3 | Patient Health Questionnaire Depression Scale (PHQ-9)

The PHQ-9 is a brief self-report measure for detecting severity of depression symptoms in a general population and was used as a measure for assessing convergent validity in the multisite and London samples. Respondents are asked to rate how bothered they have been by a range of problems over the last two weeks. There are nine items, and responses are given on a 4 point rating scale from *Not at all* (0) to *Nearly every day* (3). The PHQ-9 has high internal consistency (Cronbach's $\alpha = 0.89$), good test-retest reliability ($r = 0.84$) (Kroenke et al., 2001), and good convergent validity when correlated with the SF-20 mental health subscale ($r = 0.73$). The reliable change index for the PHQ-9 has been calculated as 5.20 (Gyani et al., 2013).

2.2.4 | Generalized Anxiety Disorder Scale (GAD-7)

The GAD-7 is a brief self-report measure to assess symptom severity of general anxiety disorder, and was used as a measure for assessing convergent validity in the multisite and London samples. As with the PHQ-9, respondents are asked to rate how bothered they have been by a range of problems over the last 2 weeks, such as "Feeling nervous, anxious or on edge." There are seven items and, as with the PHQ-9, responses are on a 4-point rating Scale from *Not at all* (0) to *Nearly every day* (3). The scale has high internal consistency (Cronbach's $\alpha = 0.92$), high test–retest reliability ($r = 0.83$), and good convergent validity against the Beck Anxiety Inventory ($r = 0.72$) (Spitzer et al., 2006). The reliable change index for the GAD-7 has been calculated as 3.53 (Gyani et al., 2013).

2.2.5 | Session Effectiveness Scale (SES)

The SES is a four-item measure that assesses patients' perceptions of the helpfulness of psychotherapy sessions (example item, "How do you feel about the session you have just completed?") (Elliott, 2000). It was used in the present study (Scottish and London samples) to assess the discriminant validity of the Goals Form, on the assumptions that patients' ratings of goal achievement would not be strongly related to their earliest ratings of session effectiveness. Internal consistency in the current samples was acceptable (Cronbach's $\alpha = 0.67$ at first rating and 0.90 at endpoint).

2.3 | Procedure

For each sample, patients were recruited either through information distributed at local public health centers, internet notices, or through the universities established counseling services. Patients who expressed an interest in the study were sent an information sheet. If they subsequently contacted the site to indicate that they were interested in participating, they were invited to an assessment interview.

The assessment interviews were conducted by the psychotherapists in the trial. At the commencement of the assessment session, the psychotherapist went over the information sheet, answered any questions, and then invited the prospective participant to sign the consent form. They were then asked to complete the CORE-OM (Scottish sample), or PHQ-9 and GAD-7 (multisite and London sample) forms, along with a demographic form. At this point, in the multisite and London samples, patients who scored 10 or more on the PHQ-9 were accepted into the study.

An assessment was then conducted with the patients, which explored their reasons for wanting psychotherapy, personal histories, current contexts, and any immediate risks. As part of this, psychotherapists were instructed to work with patients to establish goals for the psychotherapeutic work. Where this took place, this was recorded on the patient's Goals Form, along with an initial rating for each goal item. In some instances, this goal-setting process was extended in to a subsequent session (for instance, because patients were initially unsure of what they wanted from therapy). After the end of psychotherapy, patients were offered a follow-up interview, where they also completed final outcome measures (Scottish sample: Goals Form and CORE-10; multisite and London samples: Goals Form, PHQ-9, and GAD-7). This was conducted by an independent researcher (i.e., not the patient's psychotherapist), and included quantitative (multisite and London samples) and qualitative (Scottish and London samples) assessments of their experiences of using the Goals Form. This interview was attended by 47% of patients in the Scottish sample, 46.2% of patients in the multisite sample, and 60% of patients in the London sample. Where participants did not attend a follow-up interview, their last recorded score was used as an endpoint.

In the multisite and London samples, participants were offered up to 24 weeks of psychotherapy. In the Scottish sample, it was open-ended. The average number of sessions patients had was 22.2 (SD = 19.8) in the Scottish sample, 15.4 (SD = 8.2) in the multisite sample, and 18.4 (SD = 6.3) in the London sample. At the start of each session (and, in some instances, before the start), patients were asked to complete the outcome measures: Goals Form, PHQ-9, and GAD-7 (multisite and London samples); or Goals Form and CORE-10 (Scottish sample). The psychotherapists and patients then reviewed the patients' responses, including any indications of progress or deterioration, and this could then inform the direction of the psychotherapy work. At the end of each session, patients completed the SES (Scottish and London samples).

The psychotherapy being offered at each site was *pluralistic therapy* (Cooper & Dryden, 2016; Cooper & McLeod, 2007, 2011; McLeod, 2018). This is a collaborative–integrative psychotherapeutic approach in which the therapist draws on a range of established treatment methods (e.g., empathic reflections, transference interpretations, and Socratic questioning) with the aim of tailoring the intervention to the specific goals and preferences of the patient. Pluralistic psychotherapy is similar to other forms of integrative practice (particularly assimilative integrative, common factors, and eclectic approaches, Norcross & Goldfried, 2019), but puts particular emphasis on assessing and, where possible, accommodating patient preferences (Norcross & Cooper, 2021). This can be characterized as a focus on shared decision-making, in which members of the therapeutic dyad discuss—and agree on—the format and focus of the psychotherapeutic work. In the multisite and London samples, the practitioners specifically followed a manual of pluralistic therapy for depression (McLeod & Cooper, 2012). This articulated the core elements of a pluralistic approach specifically in relation to patients experiencing low mood.

The psychotherapists for each sample were either training, or qualified, counselors, psychotherapists, or counseling psychologists. They had received training in a range of established methods, primarily person-centered and psychodynamic practices; along with, at minimum, an introduction to pluralistic psychotherapy.

The collection of data for each of the three samples received ethical approval from the relevant university ethics committee.

2.4 | Analysis

Along with analyzing each data set independently, we meta-analyzed findings across samples, using a random-effects model and weighting by inverse variance. This was to establish *overall* means and 95% confidence interval, which we present in the text and the table. Meta-analysis was conducted using the software program *Comprehensive Meta-analysis* v3.0. Where meta-analysis was not feasible, we analyzed combined data from our three samples to establish "pooled" averages. As session-by-session data were available, we used within-patient, as well as between–patient, analyses. Within-patient analyses can make full use of all data, and can also present meaningful within-person variability, which is a powerful complement to between-patients analyses.

To examine whether typical quantitative characteristics of the Goals Form could be established, our analysis began with a series of descriptive explorations. For each sample, we calculated the number of goals per participant, the number of sessions for which goals were active, and baseline and endpoint aggregate session scores.

To examine the consistency of the Goals Form overtime at the between-patients level, we examined test–retest reliability of aggregate session scores at the start of psychotherapy (from assessment to the first session) and end of psychotherapy (from the last session to follow-up assessment) (where data were available). At the within-patient level, we calculated the intraclass correlation coefficient (ICC), which indicated the proportion of variance due to the individual patient as compared with the total variance. Lag-1 autocorrelation coefficients were also calculated for each sample, using multilevel modeling in R with the nlme package (R Core Team, 2008; Rizopoulos, 2006; Rosseel, 2012).

To test the convergent validity of the Goals Form, we examined the correlation of aggregate session scores against the PHQ-9 and GAD-7 (multisite and London samples), and CORE-OM/CORE-10 (Scottish sample). For our between-patients analysis, we looked at baseline and endpoint scores. For our within-patient examinations, we conducted multilevel analyses, using a random-effects model (both intercept and slope were set as random, with the exception of slope for the multisite sample), to estimate standardized regression coefficients, with the aggregate session score from the Goals Form as our dependent variable. Analysis was conducted using the nlme package in R, controlling for lag-1 autocorrelation (Pinheiro et al., 2014).

To assess discriminant validity, we examined correlations between the aggregate session score on the Goals Form and the patients' ratings of the first session using the SES.

To assess the sensitivity of the Goals Form to change in psychotherapy, we examined patients' aggregate change scores. Effect sizes (standardized mean differences, d_z) were calculated, using the standard deviation of the mean change scores as the denominator. For comparability, d_z was converted into d according to Cohen (1988) $d = d_z\sqrt{2}$ and compared against effect sizes for the nomothetic measures. The correlations of aggregate change scores between baseline and endpoint for the Goals Form and the other measures were also estimated. Session-to-session change was estimated by calculating the correlation between the aggregate session scores and session number, with random effect multilevel growth modeling controlling lag-1 autocorrelation in R with nlme package. We also calculated the difference between the aggregate session scores and the adjacent session scores using R with the *d*plyr package (Wickham et al., 2018). Session-to-session reliable change indexes (RCIs) using the standard deviation across all sessions and the lag-1 autocorrelation coefficient in each sample were also calculated.

To analyze sensitivity categorically, we assumed that (a) patients whose aggregate change score on the Goals Form was above the RCI were considered to have shown reliable improvement, (b) those who showed a reduction in aggregate change scores greater than the RCI (in a negative direction) were considered to have reliably deteriorated, and (c) others were considered not to have changed. Rates of reliable improvement, deterioration and

no change were then computed and were compared against those from the nomothetic PROMs, using Cohen's κ coefficient.

The clinical utility of the Goals Form was assessed using both quantitative and qualitative data. Quantitatively, we analyzed data from the multisite and London follow-up interviews, in which patients had been asked to rate the Goals Form (alongside other outcome and process measures) on a 1–5 scale, where 1 = *very unhelpful*, 2 = *unhelpful*, 3 = *neither*, 4 = *helpful*, 5 = *very helpful*. In the Scottish and London samples, patients had also been asked to provide qualitative descriptions of how they had experienced using the Goals Form. These answers were transcribed and subjected to thematic analysis (Braun & Clarke, 2006). We report all themes with two or more patients.

3 | RESULTS

A summary of our principal hypotheses, methods, and results is presented in Table 2.

3.1 | Descriptive analysis

3.1.1 | Number of goal items

The mean number of goal items across session (averaging the number of goals across all sessions in each sample) ranged from 3.5 (London sample) to 4.8 (Scottish sample), with an overall mean number of goal items of 4.2 (*SE* = 0.4, 95% CI [3.4, 5.0]) (Table 3). The mean number of goal items per patients (averaging the mean number of goals of each patient in each sample) were 4.7 (95% CI [4.3, 5.1]) in the Scottish sample, 4.3 (95% CI [3.9, 4.7]) in the multisite sample, and 3.4 (95% CI [3.1–3.8]) in the London sample. The overall mean was 4.1 goal items (*SE* = 0.4, 95% CI [3.4–4.9]).

3.1.2 | Length of time goal active

The mean number of sessions for which the goals were active was from 12.0 (multisite sample) to 20.3 (Scottish sample), with an overall mean of 15.9 sessions (*SE* = 2.0, 95% CI [12.1, 19.8]). The percentage of goals that were active from baseline to endpoint was 67.1% in the Scottish sample (95% CI [63.0, 71.2]), 48.8% in the multisite sample (95% CI [46.1, 51.5]), and 57.4% in the London sample (95% CI [55.3, 59.5]); with an overall mean of 57.6% (*SE* = 4.5, 95% CI [63.0–71.2]).

3.1.3 | Goal achievement at baseline

The aggregate session scores at baseline were 2.0 (London sample) and 2.4 (both Scottish and multisite sample), with an overall aggregate session score of 2.2 (*SE* = 0.1, 95% CI [2.0, 2.4]).

3.1.4 | Goal achievement at endpoint

The aggregate session scores at endpoint ranged from 4.2 (multisite sample) to 4.5 Scottish sample), with an overall aggregate session score of 4.4 (*SE* = 0.2, 95% CI [4.0, 4.7]).

TABLE 2 Summary of principal hypotheses, methods, and overall results

Hypotheses		Methods		Results
Acceptability	Rate of completed items > 90%	Between-clients level	Percentage of item completion	96.8% (*SE* = 0.01), 95% CI [92.8, 98.6]
Reliability	Test–retest reliability ($r > 0.70$)	Between-clients level	Correlation from assessment to the 1st session	0.71 (*SE* = 0.12), 95% CI [0.39, 0.88]
			Correlation from the last session to follow-up assessment	0.88 (*SE* = 0.05), 95% CI [0.73, 0.95]
	High session-to-session autocorrelations	Within-client level	Intraclass correlation coefficient (ICC)	0.53
			Lag-1 autocorrelation	0.79 (*SE* = 0.03), 95% CI [0.71, 0.85]
Validity	Moderate–to–strong negative correlations with psychological distress (−0.40 to −0.60)	Between-clients level	Correlation of GF and PHQ-9/GAD-7/CORE-10 at baseline	−0.68 to −0.14
			Correlation of GF and PHQ-9/GAD-7/CORE-10 at endpoint	−0.92 to −0.57
		Within-client level	Multilevel analyses with random-effects model, controlling lag-1 autocorrelation	−0.68 to −0.03
	Nonsignificant correlations with irrelevant variables	Between-clients level	Correlations between GF and SES at 1st session	0.14 (*SE* = 0.10), 95% CI [−0.20, 0.45]
Sensitivity to change	Large pre–post effects and statistically reliable change	Between-clients level	Pre–post change scores	2.0 (*SE* = 0.2, 95% CI [1.7, 2.3];
			Effect sizes (d)	d = 2.0 (*SE* = 0.3, 95% CI [1.5, 2.4]
			Correlations of mean change on GF and other nomothetic outcome measures	−0.40 to −0.58
	Acceptable classification consistency of reliable change between the GF and other outcome measures (Cohen's $\kappa > 0.40$)	Between-clients level	Cohen's κ coefficient of rates of reliable improvement, deterioration, and no change on GF and other outcome measures	0.21–0.69

(Continues)

TABLE 2 (Continued)

Hypotheses		Methods		Results
	Significant session-to session change	Within-client level	Difference in mean session scores across adjacent sessions	0.12 (*SE* = 0.02), 95% CI [0.08, 0.16]
			Session-to-session RCI	1.9
	Significant positive correlation between the session number and GF score	Within-client level	Correlation between the mean session scores and session number, with random effect multilevel growth modeling controlling lag-1 autocorrelation	0.12 (*SE* = 0.002), 95% CI [0.07, 0.17]
Clinical utility	On average, clients rate the GF as helpful (at least 4 on a 5-point scale, where 4 = Helpful).	Between-clients level	Mean of rating on 5-point scale	4.2 (*SD* = 1.2) and 3.9 (*SD* = 1.0)

Abbreviations: CI, confidence interval; CORE, Clinical Outcomes in Routine Evaluation Measures; GF, Goal Form; GAD-7, Generalized Anxiety Disorder Scale; PHQ-9, Patient Health Questionnaire Depression Scale; RCI, reliable change index; *SD*, standard deviation; *SE*, standard error; SES, Session Effectiveness Scale.

TABLE 3 Descriptive data for Goal Form across samples

Variable	Scottish	Multisite	London	Overall
Number of goals per client per session				
M	4.8 [4.7, 4.9]	4.3 [4.2, 4.4]	3.5 [3.4, 3.6]	4.2 [3.4, 5.0]
SD	1.1	1.1	1.3	0.4[a]
Median	5	4	3	4[b]
Mode	5	4	3	5[b]
Range	3–8	1–7	1–6	1–8[b]
N sessions	331	340	835	1506
Number of sessions goal is active				
M	20.3 [16.7, 23.9]	12.0 [10.7, 13.3]	16.4 [15.3, 17.5]	15.9
SD	16.2	7.5	7.2	2.0[a]
Median	11	12	19	15[b]
Mode	7	24	24	24[b]
Range	3–51	1–24	2–26	1–51[b]
N goals	79	121	176	376
Mean goal score at baseline per client				
M	2.4 [2.0, 2.7]	2.4 [2.0, 2.7]	2.0 [1.8, 2.3]	2.2
SD	0.7	1.0	0.8	0.1[a]
Median	2.3	2.3	2.0	2.2[b]
Range	1.4–4.2	1.2–4.8	1.0–3.7	1.0–4.8[b]
N clients	15	23	38	76
Mean goal score at endpoint per client				
M	4.5 [3.7, 5.3]	4.2 [3.7, 4.7]	4.4 [4.0, 4.8]	4.4
SD	1.6	1.4	1.5	0.2[a]
Median	5.0	4.5	4.5	4.5[b]
Range	1.6–6.6	1.5–6.4	1.0–7.0	1.0–7.0[b]
N clients	15	25	48	88

Note: Numbers in the square brackets are the 95% confidence interval of the statistical parameter in front of them.
Abbreviations: *SD*, standard deviation; *SE*, standard error.

[a]*SE*.

[b]Results of pooled sample.

3.1.5 | Variations across age and gender

The correlation between aggregate session scores at baseline and patients' age was 0.23 ($n = 15$, 95% CI [−0.32, 0.66]) for the Scottish sample, 0.02 ($n = 23$, 95% CI [−0.40, 0.43]) for the multisite sample, and 0.20 ($n = 35$, 95% CI [−0.14, 0.50]) for the London sample. The overall mean correlation was 0.15 ($SE = 0.05$) with 95% CI [−0.09, 0.38] across samples.

The correlation between aggregate session scores at endpoint and patients' age was 0.39 ($n = 15$, 95% CI [−0.11, 0.89]) for the Scottish sample, 0.06 ($n = 25$, 95% CI [−0.35, 0.47]) for the multisite sample, and 0.12 ($n = 43$, 95% CI [−0.18, 0.42]) for the London sample. The overall mean correlation was 0.16 ($SE = 0.11$) with 95% CI [−0.06, 0.37] across samples.

At baseline, the difference in aggregate session scores between males and females were 0.5 (95% CI [−0.4, 1.4], male mean = 2.7 [$SD = 1.2$], female mean = 2.2 [$SD = 0.5$], $n = 14$) in the Scottish sample, −0.1 (95% CI [−1.0, 0.8], male mean = 2.3 [$SD = 1.3$], female mean = 2.4 [$SD = 0.8$], $n = 23$) in the multisite sample, and 0.3 (95% CI [−0.4, 1.0], male mean = 2.3 [$SD = 0.8$], female mean = 2.0 [$SD = 0.9$], $n = 36$) in the London sample. The overall mean gender difference was
0.3 ($SE = 0.2$) with 95% CI [−0.2, 0.7] across samples.

At endpoint, the difference in aggregate session scores between males and females were −1.3 (95% CI [−3.4, 0.8], male mean = 3.5 [$SD = 2.2$], female mean = 4.8 [$SD = 1.4$], $n = 14$) in the Scottish sample, 0.4 (95% CI [−0.9 to 1.7], male mean = 4.5 [$SD = 1.2$], female mean = 4.1 [$SD = 1.5$], $n = 25$) in the multisite sample, and 0.7 (95% CI [−0.3 to 1.8], male mean = 5.0 [$SD = 1.3$], female mean = 4.2 [$SD = 1.6$], $n = 45$) in the London sample. The overall mean gender difference was 0.4 ($SE = 0.4$) with 95% CI [−0.5 to 1.2] across samples.

3.2 | Acceptability

Completion of individual goal items over the course of psychotherapy for each sample were 97.3% (95% CI [96.4, 98.0], Scottish sample), 93.4% (95% CI [92.1, 94.6], multisite sample), and 98.2% (95% CI [97.6, 98.6], London sample). The overall item completion across sample was 96.8% ($SE = 0.01$), with 95% confidence interval [92.8, 98.6].

3.3 | Temporal structure

3.3.1 | Test–retest correlations

From baseline to the first session, the test–retest correlation (r) for the aggregate session scores was 0.75 in the Scottish sample and 0.59 in the London sample (data for the multisite sample were not available; time interval in the London sample ranged from 1 to 11 weeks with a mode of 1 week; time interval data for the Scottish sample were not available) (Table 4). The overall mean test–retest correlation was 0.71 ($SE = 0.12$), 95% confidence intervals were [0.39, 0.88] at baseline. At endpoint, the test–retest correlation (r) for the aggregate session scores was 0.95 and 0.85 for the two samples separately (time interval in the London sample ranged from 0 to 19 weeks, with a mode of 1 week; time interval data for the Scottish sample were not available). There was an overall mean test–retest correlation of 0.88 ($n = 36$, $SE = 0.05$) with 95% CI [0.73, 0.95] at endpoint.

3.3.2 | Session-by-session reliability

The session-to-session lag-1 autocorrelations ranged from 0.74 (London sample) to 0.85 (multisite sample), with an overall mean of 0.79 across samples ($SE = 0.03$, 95% CI [0.71, 0.85]).

The ICCs ranged from 0.46 for the multisite sample to 0.55 for the London sample, with a pooled mean ICC of 0.53 across all samples.

TABLE 4 Temporal stability analysis of Goal Form scores across sessions

Variables	Scottish	Multisite	London	Overall
Test–retest reliability at baseline (assessment to first session)				
r	0.75*** [0.39, 0.91]	–	0.59 [−0.20, 0.91]	0.71 [0.39, 0.88]
n	15	–	8	23
Test–retest reliability at endpoint (last session to follow-up)				
r	0.95*** [0.69, 0.99]	–	0.85*** [0.70, 0.93]	0.88 [0.73, 0.95]
n	7	–	29	36
ICC	0.54	0.46	0.55	0.53[a]
Lag-1 autocorrelation	0.77*** [0.64, 0.86]	0.85*** [0.79, 0.89]	0.74*** [0.66, 0.80]	0.79*** [0.71, 0.85]
n	331	340	836	1507

Note: Numbers in the square brackets are the 95% confidence interval of the statistical parameter in front of them.
Abbreviation: ICC, intraclass correlation coefficient.
[a]Results of pooled samples.
***$p < 0.001$.

3.4 | Convergent validity

3.4.1 | Between patients

At baseline, the correlation between aggregate session scores and nomothetic measures ranged from −0.68 with the CORE-OM (Scottish sample) to −0.14 with the PHQ-9 (multisite sample) (Table 5). At endpoint, the correlation between the same variables ranged from −0.92 with CORE-10 scores (Scottish sample) to −0.57 with the GAD-7 (London sample).

3.4.2 | Within patient

With aggregate session scores on the Goals Form as the dependent variable, the standardized regression coefficients for the nomothetic measures ranged from −0.68 with CORE-OM/CORE-10 in the Scottish sample to −0.03 with the GAD-7 in the multisite sample.

3.5 | Discriminant validity

At baseline, the correlations between aggregate session score and patients' session ratings were −0.11 ($n = 15$, 95% CI [−0.59, 0.43]) for the Scottish sample and 0.26 ($n = 31$, 95% CI [−0.10, 0.56]) for the London sample. The overall mean correlation was 0.14 ($SE = 0.10$) with 95% CI [−0.20, 0.45] across samples.

3.6 | Sensitivity analysis

3.6.1 | Baseline to endpoint change

The aggregate change scores on the Goals Form were from 1.7 for the multisite sample, to 2.2 for the Scottish sample. The overall aggregate change score was 2.0 ($SE = 0.2$, 95% CI [1.7, 2.3]) across all patients. The effect

TABLE 5 Between-patients and within-patient convergent validity between Goal Form and other measures

		Between-clients correlation							
		Baseline		Endpoint		Baseline to endpoint change		Within-client correlation	
Measure	Sample	*r*	*n*	*r*	*n*	*r*	*n*	*β*	*n*
CORE-OM	Sco.	−0.68** [−0.88, −0.26]	15	−0.92*** [−0.98, −0.76]	14	−0.49 [−0.81, 0.06]	14	−0.68*** [−0.87, −0.49]	306
PHQ-9	Mul.	−0.14 [−0.53, 0.30]	22	−0.65** [−0.84, −0.32]	22	−0.49* [−0.76, −0.07]	21	−0.12**,[a] [−0.21, −0.03]	329
	Lon.	−0.27 [−0.55, 0.06]	36	−0.58*** [−0.74, −0.35]	47	−0.51*** [−0.69, −0.26]	48	−0.42*** [−0.51, −0.32]	833
GAD-7	Mul.	−0.29 [−0.63, −0.15]	22	−0.70** [−0.87, −0.37]	20	−0.40 [−0.72, 0.07]	19	−0.03[a] [−0.12, 0.07]	326
	Lon.	−0.45** [−0.68, −0.14]	36	−0.57*** [−0.74, −0.34]	47	−0.58*** [−0.74, −0.35]	48	−0.34*** [−0.43, −0.26]	834

Note: Within-client correlations were estimated by multilevel modeling in R with standardized outcome measures controlling lag-1 autocorrelation (nlme package in R, *df* = number of observations − number of clients − 1). Numbers in the square brackets are the 95% Confidence Interval of the statistical parameter in front of them.

Abbreviations: CORE, Clinical Outcomes in Routine Evaluation Measures; GAD-7, Generalized Anxiety Disorder Scale; Lon., London sample; Mul., multisite sample; PHQ-9, Patient Health Questionnaire Depression Scale; Sco., Scottish sample.

[a]Fixed effect model.

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

TABLE 6 Sensitivity analysis: Measuring change with the Goal Form

Variables	Scottish	Multisite	London	Overall
Pre–post change				
GF aggregate change score *M* (*SD*)	2.2 (1.4) [1.5, 2.9]	1.7 (1.5) [1.2, 2.3]	2.1 (1.5) [1.7, 2.5]	2.0 (0.2[a]) [1.7, 2.3]
ES GF (significance)	2.2*** [1.1, 3.2]	1.7*** [0.8, 2.4]	2.0*** [1.4, 2.7]	2.0*** [1.5, 2.4]
ES CORE-OM (significance)	1.0** [−1.9, 3.9]			
ES PHQ-9 (significance)		1.1*** [−1.2, 3.4]	1.0*** [−0.8, 2.8]	
ES GAD-7 (significance)		0.8** [−1.4, 3.0]	0.9*** [−0.7, 2.5]	
Session-to-session change				
Sessions (*n*)	331	299	740	1350
Difference between session at lag1 *M* (*SD*)	0.10** (0.72) [0.02, 0.18]	0.14*** (0.72) [0.06, 0.23]	0.12*** (0.86) [0.06, 0.19]	0.12*** [0.08, 0.16]
n	331	340	836	1507
Correlation with session number *r*	0.16*** [0.07, 0.25]	0.14*** [0.09, 0.18]	0.10*** [0.08, 0.12]	0.12*** [0.07, 0.17]
RCI ($p < 0.05$)	1.9	1.3	2.1	1.9[b]

Note: Difference between sessions at lag-1 was estimated in R with dplyr package. Correlation with session number was estimated by multilevel modeling with random effect controlling autocorrelation. Numbers in the square brackets are the 95% confidence interval of the statistical parameter in front of them.

Abbreviations: CORE, Clinical Outcomes in Routine Evaluation Measures; ES, effect size (d); GAD-7, Generalized Anxiety Disorder Scale; GF, Goals Form; PHQ-9, Patient Health Questionnaire Depression Scale; RCI, reliable change index; *SD*, standard deviation; *SE*, standard error.

[a]*SE*.

[b]Results of pooled samples.

**$p < 0.01$,; ***$p < 0.001$.

size of aggregate change scores for each sample was from 1.7 (multisite sample) to 2.2 (Scottish sample), giving an overall mean value of 2.0 (*SE* = 0.3, 95% CI [1.5, 2.4]) (Table 6). In comparison, the effect sizes for the nomothetic PROMs ranged from 0.8 for the GAD-7 in the multisite sample, to 1.1 for the PHQ-9 in the same sample.

The correlations between aggregate change scores on the Goals Form and change scores on the nomothetic PROMs ranged from −0.40 (Multisite sample) to −0.58 (London sample) with the GAD-7 (see Table 5).

3.6.2 | Reliable change

An RCI of 1.5 was calculated for the Goals Form, using the standard deviation and test–retest reliability of the aggregate session score at baseline of the pooled sample. Based on this figure, reliable improvement was shown by 60.0% (95% CI [34.8, 80.8]) of patients in the Scottish sample, 52.0% (95% CI [33.1, 70.4]) of patients in the multisite sample, and 54.2% (95% CI [40.1, 67.6]) of patients in the London sample. No patients, across the samples, showed reliable deterioration. The overall reliable improvement rate was 54.5% (*SE* = 5.3, 95% CI

[44.1, 64.6]). This compared against reliable improvement rates of 35.7% (95% CI [10.6, 60.8]) in the Scottish sample for the CORE-OM; reliable improvements rates in the multisite sample of 47.6% (95% CI [26.3, 69.0]) on the PHQ-9 and 57.9% (95% CI [35.7, 80.1]) on the GAD-7; and reliable improvement rates in the London sample of 54.2% (95% CI [40.1, 68.3]) on the PHQ-9 and 50% (95% CI [35.9, 64.1]) on the GAD-7.

3.6.3 | Classification consistency

The κ coefficient of reliable change was from 0.21 (with CORE) in the Scottish sample, to 0.69 (with the GAD-7) in the multisite sample (Table 7).

3.6.4 | Session-to-session change

The correlation between aggregate session scores and session number ranged from 0.10 (London sample) to 0.16 (Scottish sample), with an overall mean correlation of 0.12 (*SE* = 0.002), 95% CI [0.07, 0.17] (see Table 6).

The mean difference between sessions at lag-1 within-patient ranged from 0.10 for the Scottish sample to 0.14 for the multisite sample, with an overall mean difference of 0.12 (*SE* = 0.02, 95% CI [0.08, 0.16]) across all patients, which was significantly above 0. Using session-to-session values, the RCI of the Goals Form was from 1.3 (multisite sample) to 2.1 (London sample), giving a pooled value of 1.9 (Table 6).

3.7 | Clinical utility

3.7.1 | Quantitative

From the multisite data set, 17 ratings of the helpfulness of the Goals Form were available from patients (Figure 1). The mean rating on the 5-point scale was 4.2 (*SD* = 1.2), with a modal and median rating of 5 (*very helpful*). This compared against a mean rating of 3.7 for the PHQ-9 (*SD* = 1.0), which had been used on a session-by-session basis throughout psychotherapy; and 3.1 (*SD* = 1.0) on the Working Alliance Inventory—Short Form (Tracey & Kokotovic, 1989), which had been used at review points.

From the London data set, 35 ratings of the Goals Form were available from patients. The mean helpfulness rating was 3.9 (*SD* = 1.0), with a modal and median rating of 4 (*helpful*). This compared against a mean rating of 4.2 on the PHQ-9 (*SD* = 0.9) and 3.7 (*SD* = 1.0) on the Working Alliance Inventory—Short Form (revised) (Hatcher & Gillaspy, 2006).

TABLE 7 Classification consistency of reliable change between Goal Form and other measures (κ)

Measure	Scottish	Multisite	London
CORE	0.21 [−0.21, 0.63]		
PHQ-9		0.24 [−0.16, 0.65]	0.41 [0.15, 0.67]
GAD-7		0.69 [0.39, 0.99]	0.30 [0.07, 0.54]

Note: Numbers in the square brackets are the 95% confidence interval of the statistical parameter in front of them.
Abbbreviations: CORE, Clinical Outcomes in Routine Evaluation Measures; GAD-7, Generalized Anxiety Disorder Scale; PHQ-9, Patient Health Questionnaire Depression Scale.

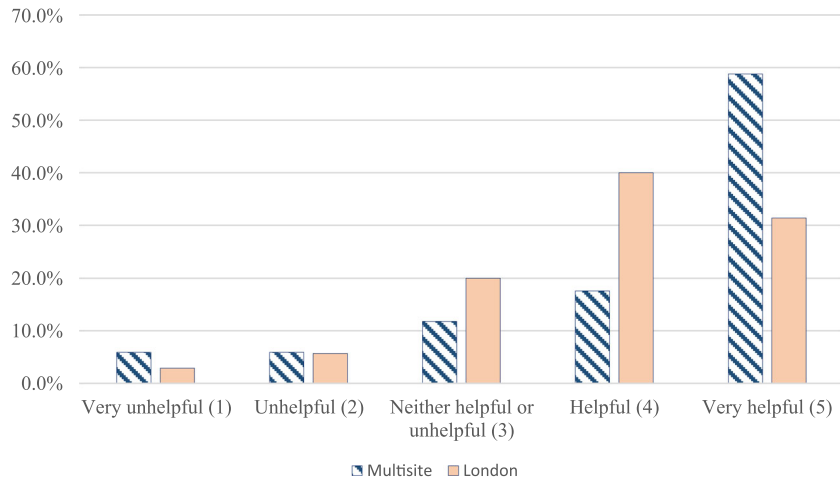

FIGURE 1 Ratings of the Goals Form note. Two midpoint ratings ("4.5" and "3.5") were rounded down for the purposes of this figure

3.7.2 | Qualitative

Two patients in the Scottish sample said that the Goals Form was of value because it helped them articulate what they wanted from therapy. For instance, "[I]t was…good to be able to have a look and see– 'Yes, this is exactly what I'm aiming at.'" Two patients also said that the use of form gave them a sense of overall progression.

In the London sample, seven patients said that the Goals Form was helpful because they liked being able to see their progress in psychotherapy: for instance, "it was nice to see the numbers going up." Six patients said that the regular use of the Goals Form was a helpful reminder of what they wanted to work on, for instance "it's kind of a prompt." Four patients said that the Goals Form was helpful because it gave focus to the therapeutic work, with one patient describing this as, "Arguably one of the most important things, actually, of the whole [psychotherapeutic] process." Closely related to this, two patients said that they found the Goals Form helped them clarify what they wanted from psychotherapy, and two patients said that they liked the personalized nature of the form.

On the critical side, four patients said that session-by-session monitoring with the Goals Form could compound feelings of failure, and three said that it should be used on a less frequent basis than every session. Three patients said that they felt routine outcome monitoring with the Goals Form could create unrealistic expectations of change. Two patients said that they got bored of using the measure.

In terms of using the Goals Form in the most effective way, two patients emphasized the need to integrate the scores in with other aspects of the psychotherapeutic work; and two patients said that it should be used flexibly, with patients able to change their goals.

4 | DISCUSSION

Our analysis established several norms for the Goals Form: mean number of goals per patient (4.2 goals), mean number of sessions for which goals were active (15.9 sessions), aggregate session scores at baseline (2.2), and aggregate session scores at endpoint (4.3). We also established a reliable change index for aggregate change scores (≥ 1.5 points). Although these norms were based on small sample sizes and may be subject to updating, there was encouraging levels of consistency across samples.

In most instances, we found that the psychometric properties of the Goals Form were acceptable. Noncompletion of individual goal items was generally low; temporal stability tended to be at, or above, our target levels (>0.70); and aggregate session scores were not significantly correlated with a theoretically independent variable (SES). We also did not find that these scores significantly varied by age or gender.

In terms of convergent validity, we found that continuous Goals Form scores, at both between-patients and within-patient levels, tended to correlate well with other indicators of psychological problems. This was large enough to indicate that the Goals Form was tapping important dimensions of psychological health, but not so large as to make the Goals Form redundant. There were, however, some exceptions to this, with nonsignificant correlations with the PHQ-9 at baseline, and a very large correlation with the CORE-OM at endpoint. In addition, rates of reliable change, based on the Goals Form, were similar to those of other nomothetic PROMs (except CORE in the Scottish sample). This is consistent with findings from studies in an adolescent population, which compared improvement rates in goal progress and on standardized symptoms or functioning measures (Krause et al., 2021). These findings show that the Goals Form, as with other goal-based I-PROMs, may capture important and unique aspects of change that are different from—but related to—those measured by nomothetic PROMs. However, the agreement between the Goals Form and nomothetic PROMs on reliable change (Cohen's κ) was not as good as we expected. This means that even, if the Goal Form and N-PROMs show moderate-to-strong quantitative correlations, it does not mean that the consistency of therapeutic outcomes, categorically, is the same. Since there has been no previous evidence in relation to this issue, we cannot judge the representativeness of our finding, but it is an important area for further inquiry.

As with other I-PROMs (e.g., Ashworth et al., 2009), we found comparatively large effect sizes for aggregate change scores on the Goals Form. This supports the hypothesis that idiographic measures may be "more capable than standardized measures of capturing relevant change for individuals" in psychological therapies (Edbrooke-Childs et al., 2015, p. 146). This potential sensitivity to important change is demonstrated, not only in terms of pre-/post- change, but also from session to session. At the within-patient level, both the mean difference scores between sessions at lag-1, and the correlation between aggregate session scores and the number of sessions, indicated that goal achievement scores increased in a consistent and relatively stable manner as psychotherapy proceeded.

We found that the convergent validity of the measure (along with its internal consistency) tended to be higher at endpoint than at the start of psychotherapy. This may suggest that, as patients progress through treatment, they develop a more coherent sense of achievement-as-a-whole, and one that is more aligned with conventional measures of psychological health. Put conversely, it may be that, at the start of psychotherapy, patients have a more fragmented and/or incoherent sense of their goals and their levels of goal achievement. This would be consistent with evidence that patients can find it difficult to formulate goals at the beginning of therapy (Di Malta et al., 2019). Closely related to this, problem-focused I-PROMs, as compared with goal-focused I-PROMs, have tended to show more internal consistency at baseline, as well as endpoint (Ashworth et al., 2005; Elliott et al., 2016). Although this may be due to the larger sample sizes of these studies, it may also be that patients have a more coherent and integrated sense—from the start of psychotherapy—of their problems, as compared with their goals. This may be because problems, as compared to goals, are more "experience near": closer to what patients are consciously thinking about and focusing on when they come in to psychotherapy.

In terms of clinical utility, responses to the Goals Form were generally encouraging. We found mean ratings in the "helpful" range, with over 70% of patients in two independent samples rating the Goals Form as either helpful or very helpful to their psychotherapeutic work. However, around 10% of patients in both samples indicated that they had not found the Goals Form helpful. In addition, one of the patients in our multisite study (2.6%), and 15 of the patients in our London sample (23.8%), did not use the Goals Form. Unfortunately, we do not have data on why this choice was made; and whether it was determined by the patient, the psychotherapist, or both. However, it adds to the body of evidence suggesting that the Goals Form may not be desired by, or appropriate for, all patients (Di Malta et al., 2019).

The principal limitation of this study is its small sample sizes, albeit within acceptable psychometric limits for our tests, and with triangulation across three independent datasets generally showing consistency of findings. To compensate for the small samples, we meta-analyzed our findings; and analyzed data at the within-patient level as well as at the between-patients level, to maximize numbers of observations. Our finding that consistency was generally greater for the former, as compared with the latter, suggests that our between-patients analyses may have been weakened by the relatively small *N*.

Another factor that may have led to an underestimation of the Goals Form's validity and reliability was our use of aggregate session scores for several of our within-patient analyses. Although this seemed the only feasible means of analyzing patients' scores over the course of treatment, as discussed earlier, the aggregate session score does not take into account changes in patients' goals. Hence, internal associations (e.g., Cronbach's α and autocorrelations) are likely to be weaker, as are associations with convergent validity measures. For future within-patient analyses, it may be prudent to only use those goals established at baseline.

We only tested the measure within the context of one form of psychotherapy. In addition, the loss of patients to endpoint interviews means that we do not know if our clinical utility findings are skewed: it may be that we have only captured the responses of the most committed, and thereby most positive, patients. Demand characteristics may also have led to more positive responses at endpoint interviews.

In terms of implications for practice, our clinical utility findings provide strong evidence that, at least for some patients, the Goals Form will be a desired and useful tool for ROM: helping to focus the psychotherapeutic work, reminding patients of their objectives, and providing evidence of progress—both on individual goals and on aggregate goal achievement. However, our evidence here also confirms that patients should not be required to use the Goals Form (Di Malta et al., 2019): some patients, at some points in time, may not experience it as clinically beneficial. Qualitative data also suggests that patients may need more than one, initial session to set goals; and may prefer to monitor their achievement on a more infrequent basis than every session. Our qualitative data also suggests that psychotherapists should be sensitive to situations where patients are not progressing on their goals, and where ROM may lead to demotivation and hopelessness. Where patients do decide to use the Goals Form, however, our validation study shows that it is a temporally robust indicator of goal achievement (both individual and aggregate form); with scores that are meaningfully aligned to other indicators of psychological wellbeing and sensitive to changes in psychotherapy.

While our between- and within-patient analyses used aggregated scores, it is unclear how clinically useful such scores may be, either at the session level or over the course of psychotherapy. Given the likely formative nature of Goals Form item, aggregated scores—or changes on aggregated scores—provide only an overall indication of progress on key personal goals. Moreover, as almost half of all goals are either established, or completed, partway through psychotherapy, the aggregate session score may not be a reliable indicator of session-by-session change. Clinically, therefore, it may be more useful for psychotherapists and patients to track goal scores at the individual item level only; though this does raise the problem of "single-item indicators," with error at the level of the construct rather than the individual scale (Diamantopoulos et al., 2008). A potential solution here could be to develop a new goal-focused I-PROM that had a series of indicators for each goal (for instance, "I feel I have attained this goal," "I am making good progress to this goal," "I am very far from this goal" (reversed)), such that a reflective measurement model could be adopted for achievement on each individual goal.

Further research into patients' experiences of using the Goals Form in different formats could be of considerable clinical benefit. For instance, how do patients experience session-by-session monitoring as compared to monthly monitoring; fixed goals versus "flexible" goals (i.e., regularly revised); and, as above, how meaningful are aggregate session and change scores to patients? Understanding the patient, therapist, treatment, or contextual factors that may mediate or moderate differences here would also be very helpful. It would also be valuable to look at the factors that may moderate or mediate patients' experiences in relation to broader questions, such as whether or not patients will find the Goals Form helpful, per se; whether patients will find goal-focused or problem-focused I-PROMs more helpful; and whether nomothetic or idiographic PROMs will be of greatest value. In relation to the

Goals Form, once the optimal conditions are identified, it would then be very useful to conduct a trial to see whether it does, indeed, add benefit to their experiences and/or outcomes. An ongoing challenge is also to develop I-PROMs that can tap unconscious, implicit "directions" as well as those goals and problems that a patient is consciously aware of (Cooper, 2019).

As with other I-PROMs (Lloyd et al., 2019; Sales et al., 2022), the use of the Goals Form for population-level purposes—such as comparison between services or treatments—raises a range of complexities. As each patient has a unique set of goals, it may be problematic to compare outcomes across services, treatments, or patient groups (Elliott et al., 2016). I-PROMs may also be more vulnerable to "gaming": with psychotherapists, for instance, setting easier treatment goals to "show" better outcomes for their service. Yet, as with Sales et al. (2022) we believe that aggregated data from the Goals Form—as an I-PROM—has the potential to contribute important additional information to population-level evaluations: indicating the extent to which patients, albeit at the broadest level, are achieving key personal goals. Establishing standardized population-level processes for indicating these effects is an important area for further research and development. Most simply, as in the present article, effect sizes for aggregate change scores could be used, or percentages of patients showing reliable change on these aggregate scores. However, the outcome metric currently being trialed with the GBO tool in Child and Adolescent Mental Health Services in England (Jacob et al., 2021) is the percentage of patients showing reliable change on at least one goal. This focus on the achievement of individual goals, rather than aggregate goal achievement, may provide a more meaningful, interpretable assessment of population-level outcomes.

Further work is also needed on specifying the measurement model underlying I-PROMs. The current, hybrid model advocated by both Sales et al. (2022) and ourselves may be utilizable, but our hope is that, for the future, more elegant and unified solutions may emerge. This will be particularly important if data from I-PROMs are to be extended into the domain of population-level analyses.

In conclusion, the Goals Form is a goal-focused I-PROM that assesses change on both individual goals and overall goal achievement. It shows psychometric properties at levels sufficient for the regular monitoring of individual patient outcomes in adult psychotherapy patients—the first measure, of its type, to be shown to do so. Item scores are temporally stable, converge with other indicators of mental health, and appear to capture relevant changes in treatment. Initial benchmarks for reliable change have been established. A significant proportion of patients find the measure helpful in their psychotherapy—though it is important to note that some do not. Further work, however, is required on the use of the measure for assessing population-level outcome: the tool has not been validated for service-or treatment-level comparisons and the underlying measurement model needs further specification. Such developments require a further program of research and, perhaps more importantly, new thinking about the meaning and validation of idiographic data.

ACKNOWLEDGEMENTS

Thanks to all patients, psychotherapists, and researchers who contributed to the collection of data.

DATA AVAILABILITY STATEMENT

Data are not available due to consent agreements at time of collection.

ORCID

Mick Cooper http://orcid.org/0000-0003-1492-2260
Dan Xu http://orcid.org/0000-0003-3572-5290

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1002/jclp.23344.

REFERENCES

Anthoine, E., Moret, L., Regnault, A., Sébille, V., & Hardouin, J.-B. (2014). Sample size used to validate a scale: A review of publications on newly-developed patient reported outcomes measures. *Health and Quality of Life Outcomes*, *12*(1), 2. https://doi.org/10.1186/s12955-014-0176-2

Ashworth, M., Evans, C., & Clement, S. (2009). Measuring psychological outcomes after cognitive behaviour therapy in primary care: a comparison between a new patient-generated measure "PSYCHLOPS" (Psychological Outcome Profiles) and "HADS" (Hospital Anxiety and Depression Scale). *Journal of Mental Health*, *18*(2), 169–177. https://doi.org/10.1080/09638230701879144

Ashworth, M., Robinson, S. I., Godfrey, E., Shepherd, M., Evans, C., Seed, P., Parmentier, H., & Tylee, A. (2005). Measuring mental health outcomes in primary care: the psychometric properties of a new patient-generated outcome measure, 'PSYCHLOPS'('psychological outcome profiles'). *Primary Care Mental Health*, *3*(4), 261–270.

Barkham, M., Bewick, B., Mullin, T., Gilbody, S., Connell, J., Cahill, J., Mellor-Clark, J., Richards, D., Unsworth, G., & Evans, C. (2013). The CORE-10: A short measure of psychological distress for routine use in the psychological therapies. *Counselling and Psychotherapy Research*, *13*(1), 3–13. https://doi.org/10.1080/14733145.2012.729069

Barkham, M., Margison, F., Leach, C., Lucock, M., Mellor-Clark, J., Evans, C., Benson, L., Connell, J., Audin, K., & McGrath, G. (2001). Service profiling and outcomes benchmarking using the CORE-OM: Toward practice-based evidence in the psychological therapies. *Journal of Consulting and Clinical Psychology*, *69*(2), 184–196. https://doi.org/10.1037//0022-006x.69.2.184

Bergman, H., Kornør, H., Nikolakopoulou, A., Hanssen-Bauer, K., Soares-Weiser, K., Tollefsen, T. K., & Bjørndal, A. (2018). Client feedback in psychological therapy for children and adolescents with mental health problems. *Cochrane Database of Systematic Reviews*, *8*(8), 011729. https://doi.org/10.1002/14651858.CD011729.pub2

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. *Psychiatric Services*, *62*(12), 1423–1429. https://doi.org/10.1176/appi.ps.002052011

Bollen, K. A. (2011). Evaluating effect, composite, and causal indicators in structural equation models. *MIS Quarterly*, *35*(2), 359–372.

Bollen, K. A., & Bauldry, S. (2011). Three Cs in measurement models: Causal indicators, composite indicators, and covariates. *Psychological Methods*, *16*(3), 265–284. https://doi.org/10.1037/a0024448

Bollen, K. A., & Diamantopoulos, A. (2017). In defense of causal-formative indicators: A minority report. *Psychological Methods*, *22*(3), 581–596. https://doi.org/10.1037/met0000056

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Cooper, M. (2015). *The Goals Form*. University of Roehampton. https://www.researchgate.net/publication/286928866_Goals_Form

Cooper, M. (2019). *Integrating counselling and psychotherapy: Directionality, synergy, and social change*. Sage.

Cooper, M., & Dryden, W. (Eds.). (2016). *Handbook of pluralistic counselling and psychotherapy*. Sage.

Cooper, M., & McLeod, J. (2007). A pluralistic framework for counselling and psychotherapy: Implications for research. *Counselling and Psychotherapy Research*, *7*(3), 135–143. https://doi.org/10.1080/14733140701566282

Cooper, M., & McLeod, J. (2011). *Pluralistic counselling and psychotherapy*. Sage.

Cooper, M., Norcross, J. C., Raymond-Barker, B., & Hogan, T. P. (2019). Psychotherapy preferences of laypersons and mental health professionals: Whose therapy is it? *Psychotherapy*, *56*, 205–216. https://doi.org/10.1037/pst0000226

Cooper, M., Wild, C., van Rijn, B., Ward, T., McLeod, J., Cassar, S., Antoniou, P., Michael, C., Michalitsi, M., & Sreenath, S. (2015). Pluralistic therapy for depression: Acceptability, outcomes and helpful aspects in a multisite study. *Counselling Psychology Review*, *30*(1), 6–20.

CORE System Trust. (2015). What's in a name (1): Scoring CORE measures. https://www.coresystemtrust.org.uk/whats-in-a-name-1-scoring-core-measures

Cox, W. M., & Klinger, E. (2021). Assessing current concerns and goals idiographically: A review of the Motivational Structure Questionnaire family of instruments. *Journal of Clinical Psychology*. https://doi.org/10.1002/jclp.23256

Di Malta, G. S., Oddli, H. W., & Cooper, M. (2019). From intention to action: A mixed methods study of clients' experiences of goal-oriented practices. *Journal of Clinical Psychology*, *75*(10), 1770–1789. https://doi.org/10.1002/jclp.22821

Diamantopoulos, A., Riefler, P., & Roth, K. P. (2008). Advancing formative measurement models. *Journal of Business Research*, *61*(12), 1203–1218. https://doi.org/10.1016/j.jbusres.2008.01.009

Diamantopoulos, A., & Winklhofer, H. M. (2001). Index construction with formative indicators: An alternative to scale development. *Journal of Marketing Research*, *38*(2), 269–277. https://doi.org/10.1509/jmkr.38.2.269.18845

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 and Adolescent Mental Health*, *20*(3), 142–148. https://doi.org/10.1111/camh.12107

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods*, *5*(2), 155–174. https://doi.org/10.1037/1082-989X.5.2.155

Elliot, A. J., & Church, M. A. (2002). Client articulated avoidance goals in the therapy context. *Journal of Counseling Psychology*, *49*(2), 243–254. https://doi.org/10.1037/0022-0167.49.2.24310.1037/0022-0167.49.2.243

Elliott, R. (2000). *The Session Effectiveness Scale*. Unpublished questionnaire. University of Toledo.

Elliott, R., Wagner, J., Sales, C. M. D., Rodgers, B., Alves, P., & Café, M. J. (2016). Psychometrics of the personal questionnaire: A client-generated outcome measure. *Psychological Assessment*, *28*(3), 263–278. https://doi.org/10.1037/pas0000174

Epton, T., Currie, S., & Armitage, C. J. (2017). Unique effects of setting goals on behavior change: Systematic review and meta-analysis. *Journal of Consulting and Clinical Psychology*, *85*(12), 1182–1198. https://doi.org/10.1037/ccp0000260

Evans, C., Connell, J., Barkham, M., Margison, F., McGRATH, G., Mellor-Clark, J., & Audin, K. (2002). Towards a standardised brief outcome measure: Psychometric properties and utility of the CORE–OM. *The British Journal of Psychiatry*, *180*(1), 51–60. https://doi.org/10.1192/bjp.180.1.51

Gyani, A., Shafran, R., Layard, R., & Clark, D. M. (2013). Enhancing recovery rates: Lessons from year one of IAPT. *Behaviour Research and Therapy*, *51*(9), 597–606. https://doi.org/10.1016/j.brat.2013.06.004

Harkin, B., Webb, T. L., Chang, B. P., Prestwich, A., Conner, M., Kellar, I., Benn, Y., & Sheeran, P. (2016). Does monitoring goal progress promote goal attainment? A meta-analysis of the experimental evidence. *Psychological Bulletin*, *142*(2), 198–229. https://doi.org/10.1037/bul0000025

Hatcher, R. L., & Gillaspy, J. A. (2006). Development and validation of a revised short version of the Working Alliance Inventory. *Psychotherapy Research*, *16*(1), 12–25. https://doi.org/10.1080/10503300500352500

Jacob, J., Edbrooke-Childs, J., Costa da Silva, L., & Law, D. (2021). Notes from the youth mental health field: Using movement towards goals as a potential indicator of service change and quality improvement. *Journal of Clinical Psychology*. https://doi.org/10.1002/jclp.23195

Jarvis, C. B., MacKenzie, S. B., & Podsakoff, P. M. (2003). A critical review of construct indicators and measurement model misspecification in marketing and consumer research. *Journal of Consumer Research*, *30*(2), 199–218. https://doi.org/10.1086/376806

Kendrick, T., El-Gohary, M., Stuart, B., Gilbody, S., Churchill, R., Aiken, L., Bhattacharya, A., Gimson, A., Brütt, A. L., de Jong, K., & Moore, M. (2016). Routine use of patient reported outcome measures (PROMs) for improving treatment of common mental health disorders in adults. *Cochrane Database of Systematic Reviews*, *7*(7), Cd011119. https://doi.org/10.1002/14651858.CD011119.pub2

Kiresuk, T. J., & Sherman, R. E. (1968). Goal attainment scaling: A general method for evaluating comprehensive community mental health programs. *Community Mental Health Journal*, *4*(6), 443–453. https://doi.org/10.1007/bf01530764

Krause, K. R., Bear, H. A., Edbrooke-Childs, J., & Wolpert, M. (2019). Review: What outcomes count? Outcomes measured for adolescent depression between 2007 and 2017. *Journal of the American Academy of Child & Adolescent Psychiatry*, *58*(1), 61–71. https://doi.org/10.1016/j.jaac.2018.07.893

Krause, K. R., Edbrooke-Childs, J., Singleton, R., & Wolpert, M. (2021). Are we comparing apples with oranges? Assessing improvement across symptoms, functioning, and goal progress for adolescent anxiety and depression. *Child Psychiatry & Human Development*. https://doi.org/10.1007/s10578-021-01149-y

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9. *Journal of General Internal Medicine*, *16*(9), 606–613.

Lambert, M. J., Whipple, J. L., & Kleinstäuber, M. (2018). Collecting and delivering progress feedback: A meta-analysis of routine outcome monitoring. *Psychotherapy*, *55*(4), 520–537. https://doi.org/10.1037/pst0000167

Law, D. (2019). *The goal-based outcome (GBO) tool: Guidance notes* (3rd ed.). MindMonkey Associates.

Little, B. R. (1983). Personal projects: A rationale and method for investigation. *Environment and Behavior*, *15*(3), 273–309. https://doi.org/10.1177/0013916583153002

Lloyd, C., Duncan, C., & Cooper, M. (2019). Goal measures for psychotherapy: A systematic review of self-report, idiographic instruments. *Clinical Psychology: Science and Practice*, *26*(3), e12281. https://doi.org/10.1111/cpsp.12281

Locke, E. A., & Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American Psychologist*, *57*(9), 705–717. https://doi.org/10.1037//0003-066x.57.9.705

Locke, E. A., Shaw, K. N., Saari, L. M., & Latham, G. P. (1981). Goal setting and task performance: 1969–1980. *Psychological Bulletin*, *90*(1), 125–152. https://doi.org/10.1037/0033-2909.90.1.125

McLeod, J. (2018). *Pluralistic therapy: Distinctive features*. Routledge.

McLeod, J., & Cooper, M. (2012). *A pluralistic approach to counselling and psychotherapy for depression: Treatment manual* (V.1 ed.). University of Abertay.

Michael, C. (2014). The Goals Form: Validating an individualised goals measure for evaluating clinical change (PsychD unpublished dissertation). Glasgow: Glasgow Caledonian University.

Moltu, C., Stefansen, J., Nøtnes, J. C., Skjølberg, Å., & Veseth, M. (2017). What are "good outcomes" in public mental health settings? A qualitative exploration of clients' and therapists' experiences. *International Journal of Mental Health Systems*, *11*(1), 12. https://doi.org/10.1186/s13033-017-0119-5

Norcross, J. C., & Cooper, M. (2021). *Personalizing psychotherapy: Assessing and accommodating client preferences*. APA.

Norcross, J. C., & Goldfried, M. R. (Eds.). (2019). *Handbook of psychotherapy integration* (3rd ed.). Oxford University.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill

Pinheiro, J., Bates, D., DebRoy, S., & Sarkar, D., R Core Team. (2014). nlme: Linear and Nonlinear Mixed Effects Models (Version R package version 3.1-117). http://CRAN.R-project.org/package=nlme

R Core Team. (2008). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org

Rizopoulos, D. (2006). Ltm: An R package for latent variable Modelling and item response theory analyses. *Journal of Statistical Software*, *17*(5), 1–25. https://doi.org/10.18637/jss.v017.i05

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Sales, C., & Alves, P. C. (2016). Patient-centered assessment in psychotherapy: A review of individualized tools. *Clinical Psychology: Science and Practice*, *23*(3), 265–283. https://doi.org/10.1111/cpsp.12162

Sales, C., Ashworth, M., Ayis, S., Barkham, M., Edbrooke-Childs, J., Faisca, J., Jacob, J., Xu, D., & Cooper, M. (2022). Idiographic patient reported outcome measures (I-PROMs) for routine outcome monitoring in psychological therapies: A position paper. *Journal of Clinical Psychology*. https://doi.org/10.1002/jclp.23319

Sales, C., Faísca, L., Ashworth, M., & Ayis, S. (2021). The psychometric properties of PSYCHLOPS, an individualized patient-reported outcome measure of personal distress. *Journal of Clinical Psychology*. https://doi.org/10.1002/jclp.23278

Spitzer, R. L., Kroenke, K., Williams, J. B., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: The GAD-7. *Archives of Internal Medicine*, *166*(10), 1092–1097. https://doi.org/10.1001/archinte.166.10.1092

Swift, J. K., Callahan, J. L., Cooper, M., & Parkin, S. R. (2018). The impact of accommodation client preferences in psychotherapy: A meta-analysis. *Journal of Clinical Psychology*, *74*(11), 1924–1937. https://doi.org/10.1002/jclp.22680

Tracey, T. J., & Kokotovic, A. M. (1989). Factor structure of the working alliance inventory. *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, *1*(3), 207–210. https://doi.org/10.1037/1040-3590.1.3.207

Tryon, G. S., Birch, S. E., & Verkuilen, J. (2018). Meta-analyses of the relation of goal consensus and collaboration to psychotherapy outcome. *Psychotherapy*, *55*(4), 372–382. https://doi.org/10.1037/pst0000170

Wickham, H., François, R., Henry, L., & Müller, K. (2018). dplyr: A grammar of data manipulation. https://CRAN.R-project.org/package=dplyr

How to cite this article: Cooper, M., & Xu, D. (2022). The Goals Form: Reliability, validity, and clinical utility of an idiographic goal-focused measure for routine outcome monitoring in psychotherapy. *Journal of Clinical Psychology*, 1–26. https://doi.org/10.1002/jclp.23344

APPENDIX: GOALS FORM

Goal 1:						
Not at all achieved						Completely achieved
1	2	3	4	5	6	7

Goal 2:						
Not at all achieved						Completely achieved
1	2	3	4	5	6	7

Goal 3:						
Not at all achieved						Completely achieved
1	2	3	4	5	6	7

Goal 4:						
Not at all achieved						Completely achieved
1	2	3	4	5	6	7

Goal 5:						
Not at all achieved						Completely achieved
1	2	3	4	5	6	7

Goal 6:						
Not at all achieved						Completely achieved
1	2	3	4	5	6	7

Goal 7:						
Not at all achieved						Completely achieved
1	2	3	4	5	6	7