Elliott, R., Wagner, J., Sales, C., Rodgers, B., Alves, P., & Café, M.J. (in press). Psychometrics of the Personal Questionnaire: A Client-Generated Outcome Measure. *Psychological Assessment*.

Psychometrics of the Personal Questionnaire: A Client-Generated Outcome Measure

Robert Elliott
University of Strathclyde
John Wagner
Dialectical Behaviour Therapy Centre of Vancouver
Célia M.D. Sales
Universidade de Évora & ISCTE - Instituto Universitário de Lisboa
Brian Rodgers
University of Queensland
P. Alves
ISCTE - Instituto Universitário de Lisboa
M. J. Café
Portuguese Association of Family and Community Therapy

Author Note

Robert Elliott, School of Psychological Science & Health, University of Strathclyde, Scotland; John Wagner, Dialectical Behaviour Therapy Centre of Vancouver, British Columbia, Canada; Célia Sales, Centro de Investigação em Educação e Psicologia, Universidade de Évora, Portugal, & Centro de Investigação e Intervenção Social, ISCTE - Instituto Universitário de Lisboa, Portugal; Brian Rodgers, School of Nursing, Midwifery and Social Work, University of Queensland, Australia; P. Alves, Centro de Investigação e Intervenção Social, ISCTE - Instituto Universitário de Lisboa, Portugal; M. J. Café, Portuguese Association of Family and Community Therapy, Lisbon, Portugal.

Brian Rodgers is now at the Department of Psychotherapy and Counselling, Auckland University of Technology, New Zealand.

This research was supported in part by funding from the University of Strathclyde and the Portuguese Foundation for Science and Technology (FCT grants PTDC/PSI-PCL/098952/2008 to Celia Sales and SFRH/BD/87308/2012 to Paula Alves). We thank Micaela Jiménez and Carlos Jiménez for statistical assistance.

Correspondence concerning this article, including Personal Questionnaire materials, should be addressed to Robert Elliott, PhD, Counselling Unit, School of Psychological Sciences and Health, University of Strathclyde, 40 George Street, Glasgow, G1 1QE, Scotland, UK. E-mail: robert.elliott@strath.ac.uk

Abstract

We present a range of evidence for the reliability and validity of data generated by the Personal Questionnaire (PQ), a client-generated individualized outcome measure, using five data sets from three countries. Overall pre-therapy mean internal

consistency (alpha) across clients was .80; within-client alphas averaged .77; clients typically had one or two items that did not vary with the other items. Analyses of temporal structure indicated high levels of between client variance (58%), moderate pre-therapy test-retest correlations (r =.57), and high session-to-session lag-1 autocorrelations (.82). Scores on the PQ provided clear evidence of convergence with a range of outcome measures (within-client r = .41). Mean pre-post effects were large (d = 1.25). The results support a revised caseness cut-off of 3.25 and a reliable change interval of 1.67. We conclude that PQ data meet criteria for evidence-based, norm-referenced measurement of client psychological distress for supporting psychotherapy practice and research.

Keywords: Outcome, measurement, individualized, psychometrics, psychotherapy
Running head: Personal Questionnaire

Psychometrics of the Personal Questionnaire: A Client-Generated Outcome Measure

Each client has a unique clinical condition, with a set of problems and presentations specific to their person and circumstances. A recurring question in outcome assessment is how to measure these unique aspects. Traditional nomothetic outcome methods using standardized measures overlook this in order to locate individuals within a larger population on general factors and norms. At the same time, existing idiographic approaches using client-generated outcome measures (CGOMs) have been criticized as both cumbersome and lacking sufficient psychometric evidence (Mintz & Kiesler, 1982). In this paper, we report psychometric analyses of an easy-to-use, simplified idiographic outcome measure, the Personal Questionnaire (PQ).

From an historical point of view, idiographic strategies in psychology were first espoused by Gordon Allport (1937), who later wrote, "as long as psychology deals with universals and not with particulars, it won't deal with much" (Allport, 1960, p.146). Pascal and Zax (1956) were among the first to use CGOMs when they defined individual behavioral outcome criteria for 30 psychiatric inpatients using clinical records. Kiesler (1966) emphasized the need to consider the diversity of clients, therapists, and treatments and Rickard (1965) used the term "tailored" to refer to assessment criteria chosen on a case-by-case basis.

CGOMs have grown in popularity in the last two decades. A review of 116 psychotherapy outcome studies published in the *Journal of Consulting and Clinical Psychology* between 1986 and 1991 revealed that they were almost never used (Lambert & McRoberts, 1993). In contrast, a recent review (Sales & Alves, 2014) reported the use of CGOMs in many research contexts, from naturalistic studies to experimental designs (e.g., Alves, Sales & Ashworth, 2013; Elliott et al., 2009; McLeod, Elliott & Rodgers, 2012). This review also identified three main CGOMs in use today: Goal Attainment Scaling (GAS, Kiresuk & Sherman, 1968), PSYCHLOPS (Ashworth et al., 2004), and the simplified version of the Personal Questionnaire (Elliott, Mack & Shapiro, 1999). The PQ was found to be the most popular CGOM, used in 11 published studies (Sales & Alves, 2014).

Despite the growing popularity and use of CGOMs, they have been viewed with some skepticism. In reviewing them, Mintz and Kiesler (1982) noted that many studies utilizing these techniques have not specified the manner of eliciting items or calculating scores from one study to the next. A second problem is the limited

psychometric data for these measures, including empirical evidence for their validity. For example, while the GAS has been widely used, it lacks psychometric research. (On the other hand, Ashworth and collaborators, 2007, studied provided some limited psychometric analyses for PSYCHLOPS.)

The original Personal Questionnaire, developed by M.B. Shapiro (1961), was an individualized, client-generated self-report measure designed to measure changes in specific psychological difficulties, in a way that allowed for comparison between different clients and different aspects of a given client's problems. Shapiro's original method proved cumbersome, however, and so it was later modified by Shapiro and others (McPherson & LeGassicke, 1965; Phillips, 1986; Shapiro, 1969).

In this article, we present detailed psychometric analysis of data generated by a simplified version of the PQ. In this version, a clinician (intake worker, therapist, researcher) helps the client through a process of developing a list of approximately ten problem statements describing in their own words what they want to work on in treatment; the client then rates these problems on a 7 point scale. The process of constructing the list of PQ problem statements generally takes about 30 minutes and can also be included within an intake or first or second therapy session. Once the PQ is constructed, clients typically complete the PQ at the beginning of each therapy session, generally taking less than a minute to do so.

The current version of the PQ has recently been integrated with standardized outcome measures in a variety of contexts of psychotherapy research, namely, in hermeneutic single-scale efficacy studies (e.g. Carvalho, Faustino, Nascimento & Sales, 2008; Elliott et al., 2009; MacLeod et al., 2012); methodologically pluralistic approaches to client change processes in psychotherapy (e.g. Klein & Elliott, 2006); randomized clinical trials to study the efficacy of psychotherapy (e.g. Barkham, Shapiro & Firth-Cozens, 1989; Vieira, Torres & Moita, 2011); and multiple case-study designs (e.g. Grafanaki & McLeod, 1999). Several of these studies have been conducted in the context of practice-based research networks, such as the International Group for Personalizing Health Assessment (IPHA, Sales, Alves, Evans & Elliott, 2014).

When it comes to the clinical utility of the PQ, a small study by Sales and collaborators (2007) reported that nearly 60% of therapists surveyed used the PQ for clinical and research purposes. These therapists relied on the PQ for several clinical tasks, such as preparation for sessions (92% of respondents) and for post-session discussions (75%). Amongst the advantages of the PQ, therapists reported usefulness for session-to-session outcome monitoring (38%), enhancing knowledge of client specific complaints (33%), and clinical decision making (21%). Disadvantages included the need for extra time and human resources (14%), overload of information about clients (24%) and the risk of an excessive focus on the client's point of view (48%). However, most therapists surveyed reported interest in integrating the PQ into their routine clinical practice (92%). In order to assist therapists in the routine use of the PQ, it has been integrated in a personalized outcome management web-based system, the Individualized Patient Progress System (IPPS, Sales & Alves, 2012; Sales et al., 2014).

Despite its clinical appeal and increasing use, there is very little previous published psychometric research on any version of the PQ, and almost all of it has used earlier versions. Phillips (1986) in extensively reviewing early versions of the PQ focused entirely on the statistical significance of various measures of internal consistency, but reported no standard parameters such as correlations or alpha. Egan, Miller and McLellan (1998) reported reliability and validity data but used a

standardized list of anxiety items. Using an earlier version of the PQ, Barkham, Shapiro, and Firth-Cozens (1989) reported evidence for reasonable convergence with data generated from two other symptom measures, the SCL-90 (Derogatis, 1983; r =.45) and the Present State Examination Scale (PSE; Wing, Cooper, & Sartorius, 1974; r=.41). Similarly, with the same data set, Barkham, Stiles, and Shapiro (1993) found that client mean improvement across treatment on the PQ correlated with mean change on the Beck Depression Inventory (r=.44), the PSE (r=.43), and the SCL-90 (r=.37). Barkham et al. (1996) did look at the current version of the PQ, but reported only pre-therapy test-retest correlations for four separate content groupings of PQ items, ranging from .49 (mood) to .56 (symptoms), with corresponding item-level minimum reliable change values from 2.47 to 1.89.

Thus, in spite of their intuitive appeal, it seems likely that researchers have not really viewed individualized outcome measures in psychometric terms, which would put them outside the domain of evidence-based assessment (Hunsley & Mash, 2007), a strong argument against their continued use. To address this situation, we undertook a detailed examination of the psychometric properties of the PQ data. In this paper, we present and integrate data from five different data sets from three countries (USA, Scotland, Portugal), including both general outpatient and specialized client populations (depressed, socially anxious). Specifically, we examine the following sets of psychometric propositions or hypotheses in order to generate a network of evidence regarding the use of PQ scores in psychotherapy outcome assessment: (1) *Normative:* Typical quantitative characteristics of PQ scores can be established, including number of items, initial severity and duration of problems. (2) *Internal structure:* PQ scores will show substantial levels of internal consistency (alpha $\geq$.7), will have relatively few inconsistent items ($\leq$2), and will be generally (>50%) unidimensional (which would support the use of a single index of weekly client problem distress),. (3) *Temporal structure:* PQ scores over time will be strongly consistent, showing large pre-therapy test-retest correlations, substantial pre-post correlations, and high levels of statistical nonindependence in the form of session-to-session autocorrelations and variance accounted for by clients. (4) *Construct validity:* PQ scores will show moderate to strong correlations (correlations in the .4 - .6 range) with standardized outcome measures of psychological distress (general distress, specific symptoms, self-relationship, psychological functioning), but will not correlate so strongly as to indicate redundancy with these (correlations > .7); in addition, for the small number of discriminant validity associations assessed, we expected less than strong relationships (correlations < . 4). (5) *Sensitivity to change:* PQ scores will be able to detect client change session-to-session and over the course of psychotherapy, showing large pre-post effects and statistically reliable change. Optimal clinical cut-off and reliable change threshold values can be established for PQ scores.

Method

We used a five-sample replication design to assess a wide range of psychometric parameters, then took a meta-analytic approach in order to derive overall estimates of these parameters. The five samples come from five different psychotherapy outcome studies carried out between 1986 and 2013 by the various combinations of the authors. Table 1 provides an overview of the methods used across the five samples, including number, gender, age, and ethnicity of clients recruited, type of therapy offered, number of therapists, number of therapy sessions offered and delivered, number of comparator instruments, and years of data collection. Because the PQ was used across all five samples, it will be described first.

Personal Questionnaire

As noted, the PQ is a client-generated individualized outcome measure designed to measure changes in individualized psychological difficulties in a consistent manner (for procedure manual and blank forms, see Elliott et al., 1999; Portuguese version: Sales et al., 2007). Items were first elicited from clients using a simple, open-ended *Problem Description Form*, which asked them to describe the problems that led them to seek treatment and that they wanted help with in therapy. A trained interviewer (usually an intake worker or researcher) then reviewed this list, transferring the problems onto individual note cards. In this process, the interviewer asked whether the client wanted to include any problems for each of five topic areas (if not already given): symptoms, mood, specific performance, relationships, and self-esteem; they then helped the client separate complex statements, clarified ambiguous statements, and encouraged the client to discard redundant statements, in order to arrive at a list of approximately ten simple, nonredundant problem statements. After the list of problems was finalized, the interviewer asked the client to rank order them from most important to least important. The client was then instructed to "rate each of the following problems according to how much it has bothered you during the past seven days, including today," using a 7-point anchored scale (1: "Not At All"; 2: "Very Little"; 3: "Little"; 4: "Moderately"; 5: "Considerably"; 6: "Very Considerably"; 7: "Maximum Possible"). Finally, they were asked to rate *problem duration*, also on a 7-point anchored scale (1: "less than 1 month"; 2: "1 - 5 months"; 3: "6 - 11 months"; 4: 1 - 2 years; 5: 3 - 5 years; 6: "6 - 10 years"; 7: "more than 10 years"). (This last procedure was not done for the USA depression dataset.) Afterwards, the client's PQ was typed up, leaving space for them to note any additional difficulties that they might subsequently experience. On subsequent administrations, clients rated severity (for the past week) only.

(1) USA Depression Sample

Participants. As part of an open clinical trial of a new treatment for depression, 48 clients were primarily recruited through advertisements in local newspapers (see also Table 1 for participant information). Six percent were Hispanic-American, 2% were African-American and the rest were European-American. Using the Diagnostic Interview Schedule (Robins, Helzer, Croughan & Ratliff, 1981) administered by trained research staff, all of the clients either met DSM-III diagnostic criteria for current major depressive disorder or were diagnosed with related affective disorders, either minor depression or atypical bipolar disorder (i.e., current major depressive episode plus a history of hypomanic symptoms). Clients were excluded for a variety of reasons (previous psychiatric hospitalization or bipolar, schizophrenic, or antisocial personality disorders; recent substance abuse or eating disorder; recent therapy or counseling; or active suicidal state). Ten therapists were involved in the study: One was a licensed clinical psychologist; one was a postdoctoral fellow; the rest were graduate students in clinical psychology.

Procedure. Participants completed several measures prior to beginning treatment and before and after each therapy session. Clients were offered up to 20 sessions of an early version of emotion-focused therapy (EFT; Elliott, Watson, Goldman & Greenberg, 2004). Of the treatments, 27 clients completed 12 sessions or more, 17 involving clients with major depressive disorder and fully trained therapists and 10 involving training clients, who had related affective disorders and were seen by therapists in training.

Measures. A battery of measures was used to examine change and to provide evidence for the convergent validity of PQ scores; however, only data from the *Symptom Checklist-90-R* (SCL-90-R; Derogatis, 1983) were complete enough to be reported here. The SCL-90-R is a standard self-report measure of psychiatric symptoms for which extensive psychometric data are available, with higher scores indicating greater distress or dysfunction. The Global Symptom Index (GSI, mean of all 90 items) was used as a measure of general clinical distress. (Internal alpha for this sample was .97.)

(2) USA General Outpatient Sample

Participants. Sixty-four clients were primarily recruited through advertisements in local newspapers offering up to 40 free sessions of experiential psychotherapy for personal or interpersonal difficulties as part of a research study and provided PQ data for at least one session (see also Table 1). Ten per cent gave their ethnicity as Hispanic-American or African-American and the rest were European-American. Admission criteria were liberal and clients were seen for a variety of DSM-IV Axis I and Axis II disorders. A small number of clients were excluded, however, because they were actively suicidal, already receiving counseling services elsewhere, or were diagnosed with acute primary substance or alcohol dependence. The most common diagnoses (assessed by a trained researcher) were affective (84%) or anxiety (53%) disorders; 44% had Axis II disorders (on the SCID-II; First, Spitzer, Gibbon & Williams, 1997; multiple diagnoses were common). Twenty-eight listed a current medication for a psychological condition. One therapist was a licensed clinical psychologist and the rest were graduate students in clinical psychology.

Procedure. Participants completed a variety of self-report measures prior to beginning therapy. They also completed the PQ before starting each therapy session. Treatment outcome was assessed every 10 sessions via self-report measures. Clients received anywhere from 1 to 63 sessions of Emotion-focused therapy.

Measures. Clients completed the following (higher scores indicate greater distress or dysfunction unless otherwise stated): (a) *Harter Global Self-Worth Scale* (Harter) is a 6-item self-report subscale of Messer and Harter's (1986) Adult Self-Perception Profile, used to measure global feelings of self-worth (internal alpha for this sample was .88). Higher scores indicate greater levels of global self-worth. (b) *Inventory of Interpersonal Problems – 26* (IIP-26; Horowitz, Rosenberg, Baer, Ureno & Villaseñor, 1988) is a self-report measure developed to assess distress about interpersonal difficulties (e.g., intimacy, assertiveness). The 26-item short form was developed by Maling, Gurtman and Howard (1995) (internal and test-retest reliabilities range from .80 to .98; for this sample internal alpha was .91). (c) *NEO Five-Factor Inventory* (NEO-FFI; Costa & McCrae, 1992) is a 60-item self-report questionnaire that provides a brief, comprehensive measure of the five domains of personality; here we focus on results for the Neuroticism subscale but also mention results for the 4 other subscales (for which higher scores indicate more Extraversion, Openness to experience, Agreeableness or Conscientiousness). (For this sample, internal alphas were Neuroticism, .85; Extraversion, .84; Openness to Experience, .77; Agreeableness, .68; and Conscientiousness, .87.) (d) *Clinical Outcome Routine Evaluation – Outcome Measure* (CORE-OM; Mellor-Clark et al., 1998; Evans et al., 2002) is a standardized 34-item self-report measure of psychological distress using a 5-point anchored frequency scale (ranging from 0: "not at all" to 5: "most or all of the time"), with a one week time frame; extensive psychometric data are available (e.g., Evans et al., 2002). (Internal alpha for this sample was .95.) (e) Global Assessment of

Functioning (GAF; American Psychiatric Association, 1993) ratings were completed by therapists at beginning and end of therapy. (f) Clients also completed the SCL-90-R, also used in USA Depression sample (internal alpha for this sample was .97).

(3) Portugal Outpatient Sample

***Participants*.** A convenience sample following a practice-based research approach was constructed by inviting three free psychotherapy services of varying lengths (University Counselling Service of the University of Madeira, Department of Psychiatry of São João Hospital and Psychotherapy Service of the Higher Institute for Applied Psychology) and two private practice psychodrama group therapists (who ran groups of varying lengths) to join an on-line practice-based psychotherapy research network, "Psychotherapy Research Portugal" and to pilot a new outcome management system IPPS (Sales & Alves, 2012; Sales et al., 2014). A total of six therapists participated (all female), and 72 patients were recruited. Most of the clients (71%) had applied for individual therapy while the rest were beginning psychodrama (29%). The majority of patients in this sample (76%) lacked a formal diagnosis; of the clients with formal diagnoses (assigned by their therapists), 26% had anxiety or panic disorders. Most were single or divorced (61%) and had some university education (74%). (See also Table 1.)

***Procedure*.** Participants in this pilot study were offered use of the measures available in the IPPS system at the pre-treatment stage and subsequent sessions. After consenting to participate in the study, all therapists were provided with a brief training session and manuals on how to use the system and its measures. All new clients were then invited by therapists to take part in the study before starting treatment. Upon consent, the PQ interview took place, to create the client-generated list of items, together with the CORE-OM and PHQ-9. The PQ interviews were conducted by six licensed clinical psychologists. Subsequently, the PQ was administered individually in paper form to patients before each session, either by their therapists or another member of the clinical team. After the sessions, the client responses were entered into the IPPS, and used for monitoring the patient's progress.

Measures. This study used Portuguese translations of two measures to examine change and provide evidence for the convergent validity of PQ scores; in both cases higher scores indicated greater distress or dysfunction: (a) *Clinical Outcome Routine Evaluation – Outcome Measure* (CORE-OM; Portuguese version: Sales, Moleiro, Evans & Alves, 2012), used also in sample 2 above. (Internal consistency for this sample: .93.) (b) *Patient Health Questionnaire - 9* (PHQ-9; Kroenke, Spitzer & Williams, 2001) is a 9-item self-report measure evaluating the DSM-IV criteria for depression on a 4-point anchored scale (from 0: "not at all" to 3: "nearly every day"); evidence for good reliability and validity have been reported for PHQ-9 scores. (Internal consistency for this sample: .87.)

(4) Scotland Social Anxiety Sample

***Participants*.** Clients were primarily recruited through advertisements in local supermarkets or referred by local mental health agencies for a study offering up to 20 sessions of free humanistic psychotherapy for social anxiety; sixty four clients provided PQ data at screening or least one session (see also Table 1). To be accepted into the study clients had to see themselves as having a problem with social anxiety and to meet DSM-IV criteria for social anxiety (assessed by a trained researcher using the SCID; First, Spitzer, Gibbon & Williams, 2002), judged as their main presenting problem. In addition to social anxiety, other common diagnoses were depression and

generalized anxiety. Thirty percent listed a current medication for a psychological condition. Fifteen therapists were involved in the study, 11 female and four male. Ten were postgraduate diploma or MSc level counselors; the rest were PhD level in counselling (2), counseling psychology (2) or clinical psychology (1).

Procedure. Clients were offered up to 20 free sessions of either person-centered or emotion-focused therapy for social anxiety and completed a variety of self-report measures prior to beginning therapy. They completed the PQ at screening, at the beginning of each therapy session, at mid- and post-therapy, and at 6- and 18-month follow-ups.

Measures. In addition to the PQ, several other outcome measures were used; except where noted higher scores indicated greater client distress or dysfunction: (a) The *CORE-OM*, as in samples 2 and 3 (internal alpha for this sample: .95). (b) The *Social Phobia Inventory* (SPIN; Connor et al., 2000) is an eleven-item problem specific measure of social anxiety symptoms supported by evidence for good reliability and validity (internal alpha for this sample: .93). (c) The 26-item version of the *Inventory of Interpersonal Problems* (IIP-26) is an interpersonal problem distress measure, also used in sample 2 (internal consistency for this sample: .90). (d) The *Strathclyde Inventory* (SI; Freire, 2007) is an experimental person-centred outcome measure assessing a single 31-item dimension of congruence/fluidity vs incongruence/structure-boundness (internal alpha: .93), scored in the direction of higher client functioning. (e) The *Self-Relationship Scale* (SRQ; Faur & Elliott, 2007) is an experimental measure used to measure the client's relationship to self; two subscales were used: Self-Attack (7 items; internal alpha: .79) and Self-Affiliation (scored in the direction of higher client functioning; 10 items; internal alpha: .94).

(5) Scotland Outpatient Sample

Participants. Clients were primarily recruited through advertisements in local supermarkets or referred by local mental health agencies for a study that offered up to 40 sessions of free person-centered/experiential psychotherapy for "personal and interpersonal difficulties." PQs were constructed at intake for 207 clients, and 188 (91%) provided PQ data for at least one session. Admission criteria were liberal and in keeping with the philosophy of the agency clients were not formally diagnosed; however, the most common presenting problems were interpersonal and self-concept issues; other common issues were dealing with emotions, life functioning problems, depression and anxiety. A small number of clients were excluded because they were actively suicidal, already receiving counseling services, and/or were diagnosed with severe substance abuse or current domestic violence. Thirty-eight percent listed a current medication for a psychological condition. (See also Table 1)

Procedure. Participants completed several self-report outcome measures at the beginning and end of psychotherapy, every 10 sessions and at optional 6- and 18-month follow-ups. They completed the PQ at intake, before starting each therapy session, and at the same time as the other outcome measures. Clients received from 0 to 44 sessions of PCE therapy. The study took place in a university-based research/training clinic and used predominantly student therapists who were learning PCE therapy. Thirty-three therapists were involved in the study, 27 female and 6 male; 15 were diploma level student counselors; 16 were counseling psychology doctoral students; two were doctoral level practitioners in either counselling or counseling psychology.

Measures. This sample used a subset of the measures used in sample 4: In addition to the PQ, these included the *CORE-OM* (internal alpha for this sample: .95) and the *Strathclyde Inventory* (SI; internal alpha for this sample: .96).

Analysis Approach

Weekly PQ scores are multi-level data, and also have varying numbers of items for different clients and even for the same client on different weeks. Because the PQ is meant to be used for repeated measurement over time in longitudinal or case study research (cf. Elliott et al., 2006; Accurso, Hawley & Garland, 2013), it is essential to assess the psychometric characteristics of scores at within-client as well as between-client levels, and to address issues of nonindependence. The multilevel nature of the data thus required a complex analysis strategy: Where possible (e.g., convergent validity analyses), we used multilevel analyses; however, in other cases (inter-item structure analyses) the complexity of the data prevented the use of more sophisticated multilevel approaches. In addition, our main focus in addressing our research questions was on patterns across the five samples; therefore, we adopted an integrative, meta-analytic approach to presenting the results of this study, organized topically rather than by sample. In calculating overall cross-sample values for psychometric parameters, we used a random effects model, weighting effects by inverse variance, following Borenstein, Hedges, Higgins and Rothstein (2009). However, where relevant, substantial deviations of samples from the overall result will be noted. (See Table 2 for a summary overview of the five sets of analyses.)

Results

Descriptive Analyses

In order to begin the process of establishing normative data about the PQ, we carried out descriptive analyses of number of PQ items, severity levels at beginning of therapy and throughout, and duration of PQ problems. (Overall values were weighted by sample size.)

Number of PQ Items. Because the number of PQ items created and rated was largely determined by the client and could even vary slightly from session to session, as clients added items or left particular items blank, we calculated the number of items rated across all 7107 sessions (see Table 3). Across samples, the weighted mean number of items rated was around 10 (mean = 9.5; sd = 2.8), with only the Portuguese sample varying substantially (mean = 5.1; sd = 2.2), probably due to differences in the administration of the PQ scale construction interview.

Severity ratings. The most useful normative or baseline value is the initial mean value for cross-client distress at screening or the beginning of therapy (n = 427 clients; mean = 5.04; sd = .93), which corresponds to "considerably" distressed. This value can be employed for interpreting client initial PQ scores, for example, by using it to establish a *caseness* cut-off or threshold value according to Jacobson criterion a (Jacobson & Truax, 1991), defined as pre-treatment clinical population mean minus 2 sd (i.e., < 5% probability of belonging to a normative clinical population). Applying this criterion to our data yields a clinical cut-off value of 3.18, which can be rounded to 3.25, the nearest quarter point.

Prior duration of problems. Assessing clients' perceptions of the prior duration of their PQ problems is done during the PQ creation process and was only added part-way through collection of the USA outpatient sample to provide a retrospective temporal baseline against which to measure client change in therapy. (This is particularly useful for systematic case study research.) Overall, clients (n =

352) rated their problems as having bothered them at roughly the same level for 3 - 5 years, i.e., a mean rating of 4.98 points (sd = 1.36). Socially anxious clients reported the longest problem duration (corresponding to 5 – 10 years); clients in the Portuguese samples reported the shortest duration of problems (roughly 1 – 2 years). (See Table 3.)

Internal Reliability Analyses

In order to assess internal consistency of item scores we (a) looked separately at both between-client and within-client levels, and (b) set minimum numbers of observations (either clients or sessions within clients) in order to enhance the stability of estimates. At the within-client level, we also examined internal item structure of PQ severity ratings in various ways, including internal consistency (alpha), number of inconsistent items, and number of underlying dimensions or factors.

Between-clients. To examine internal consistency at the between-client level for each sample, pre-therapy PQ scores were analyzed across clients repeatedly for 2 to 13 items (as long as n was at least 20 clients, a somewhat arbitrary value selected in order to increase stability of estimates). In other words, we used a resampling strategy (Good, 2006) in which we started with 2 items (which had the largest client sample) and gradually increased the number of items until the number of clients with at least that many PQ items fell below 20. The ranges of items and sample sizes for these repeated analyses are reported in Table 4, as are their mean and standard deviation summary values. Overall mean alpha (weighted by inverse variance) across samples was .80 (standard error: .03), with the lowest value for the USA depression sample (.71) and the highest value for the USA outpatient sample (.87).

Within-clients. In order to obtain reasonably stable estimates of inter-item internal reliability at the within-client level, PQ data from each client were separately calculated using the maximum number of items for which there was data from at least 10 sessions (a standard block of sessions in several of the samples; see Table 5). Although there was substantial variability between clients within samples, mean alphas were quite consistent across samples, with an overall alpha of .77 (n = 236). In general, data from 77% of clients had alphas of at least .70, which we used as the level of sufficient internal consistency. The Portuguese sample had the lowest level of adequate alphas (66%), probably because of the smaller mean number of items; data from clients in the Scottish social anxiety sample indicated most consistent item ratings (86%). Unsurprisingly, given natural clinical complexity and how PQ items are constructed for nonredundancy, clients typically had 1 or 2 items that were not internally consistent with the rest of the PQ items, defined by corrected item-total correlations of less than .3 (overall mean number of inconsistent items =1.7; sd = 2.3). The Portuguese sample had the smallest number of inconsistent items (mean = 1.0; sd = 1.4); the USA depression sample had the largest number (mean = 2.0; sd = 2.6).

Dimensionality

The existence of items inconsistent with the rest of the scale raises the possibility that, for these clients, the PQ is assessing multiple dimensions of psychological distress. Thus, we again used a resampling strategy in which we analyzed PQs for each client with varying numbers of items (i.e., 2 – 12) using principal components analysis (PCA; SPSS Factor procedure; eigenvalue = 1 criterion; 2 to 12 items; casewise deletion of data); for each client we selected the solution with the maximum number of items such that there was data from at least 10

time points (see Table 5). We used PCA and eigenvalue = 1, rather than principal axis analyses and more conservative criteria because PCA is more robust with small numbers and in order to work almost certainly yielded an overestimate of the number of actual factors (Gorsuch, 1997).

Across client samples, the mean number of factors extracted was 2.4 (sd = 1.1), a value that was consistent across all the samples except for the Portuguese sample (mean = 1.8; sd = 1.4), likely due to the smaller number of PQ items there. Overall, one-factor solutions fell significantly below our expectations in that they were obtained for only 23% of clients, ranging from 10% for the two USA samples to 41% for the Portuguese sample. (Exploratory analyses found that number of PQ items and variability [standard deviations] across PQ item mean levels within clients both predicted number of factors in several of the data sets, consistent with the likelihood that too many factors were extracted in the PCAs.)

Temporal Structure

Temporal consistency within clients is a key issue in tracking outcome over time, especially in case study research and to assess typical levels of statistical nonindependence in weekly PQ tracking. We thus undertook a series of analyses of scale-level consistencies in PQ scores over time, including test-retest correlations and various time series parameters, as shown in Table 6, using meta-analytic methods (random effects model, weighted by inverse of sample variance) to generate overall values across samples.

Test-retest correlations. To obtain classic test-retest reliability estimates (which can be used for calculating reliable change index values; Jacobson & Truax, 1991), we used data from the four samples in which the PQ was administered both at intake and before session 1 of therapy, making it possible to evaluate test-retest consistency in the absence of therapy. These ranged from a correlation of .39 for the USA outpatient sample to .73 for the Portuguese sample, with an overall value of .57 (n = 353; 95% confidence interval: .43 to .68. Mean days between intake and session 1 administrations of the PQ varied between 13 (Portuguese sample) and 48 (Scottish Social Anxiety sample), with an overall mean of 34 days. Pre-post correlations over therapy were more consistent and, unsurprisingly, somewhat lower, with an overall value of .41 (n = 345; 95% CI: .31 to .49).

Time series parameters. Weekly administration of the PQ creates time series data with potentially complex mathematical structures of nonindependence; these need to be understood in order to construct the best methods of analysis. To assess the overall level of statistical nonindependence in the data sets, we calculated eta-squared values for the total *between-client variance* in each data set (see Table 6 row f). Overall variance attributable to clients (eta-squared; random effects model) was .58 (n = 7107; 95% CI: .52 to .63); the only sample whose eta-squared value fell outside this confidence interval was the USA outpatient sample, which was higher (eta-squared = .68). Next, we looked at session-to-session (lag 1) within-client *autocorrelation*s, in which successive scores (here pooled but with breaks between clients) were correlated. These values (see Table 6, row c) also pick up both between-client variance and secular trend, were also quite large and generally consistent across samples, with an overall weighted correlation of .82 (n = 6412; CI: .78 to .85); only the Portuguese sample correlation fell outside (below) the confidence interval (r = .73).

After that, we decomposed the temporal structure into components, roughly following key parameters in ARIMA modeling (Glass, Wilson, & Gottman, 1975).

First, we looked at *secular trend* (also referred to as *nonstationarity*) over the course of therapy, which was assessed by correlating weekly PQ scores with session number, using data up to session 20, for comparability across samples (see Table 6, row d). There was a fair amount of variability across samples, with an overall weighted mean r of -.28 (n = 5630; CI: -.4 to -.15), a medium effect size indicating that PQ scores were in general moderately nonstationary. However, the USA outpatient sample showed less consistent improvement associated with session number (r = -.09), while both the 20-session time-limited samples focused on particular client presenting problems (USA depression and Scotland social anxiety clients) showed higher levels of nonstationarity (r = -.45 and -.44 respectively). This pointed to the value of working with the *difference* between successive PQ scores in order to control for nonstationarity.

Second, we assessed the *autoregressive* (correlated session-to-session error) time series component by differencing successive PQ scores within clients to eliminate secular trend and between-client differences in level of PQ scores and assessing for autocorrelation in the differenced scores (see Table 6, row e). Overall, a substantial weighted autoregressive component was clearly present: -.37 (n = 5630; CI: -.40 to -.35), with only the USA depression sample value falling slightly outside the confidence interval (r = -.42). Thus, the general temporal structure of PQ time series data here was both *nonstationary* and *autoregressive*.

Convergence with Standardized Psychotherapy Outcome Measures

In order to evaluate the scale-level convergence between standardized outcome measures and PQ scores when used over the course of therapy to assess outcome, all client outcome assessments were utilized in the analyses, with multiple data points over time per client for four of the data sets (i.e., screening, pre-, post- and one or more mid-therapy assessments) (The exception was the Portuguese data for which only pretherapy data were available for the standardized measures.) Sampling from different points in therapy is important in assessing outcome measures in order to pick up change over time but results in multilevel data sets that need to be deconstructed (e.g., Rush & Hofer, 2014). Accordingly, we first analyzed *between-client* correlations for the means of all assessment points for each client. Then, we separately examined *within-client* correlations, controlling for client differences in mean level (intercept) on measures thus assessing convergence over therapy. We carried out the latter analyses with the lme4 package (Bates, Maechler, Bolker & Walker, 2014) within R (R Core Team, 2012), using a random intercepts model (i.e., treating client mean scores as random effects). Significance tests used degrees of freedom corrected for number of clients (df = n assessments – k clients – 1; Snijders & Bosker, 2011).

Finally, a meta-analytic approach was taken to combining results across data sets using a random effects model (weighting by inverse error), carried out separately for between-client and within-client correlations; measures. Although the five data sets used a wide range of different outcome measures, these fell naturally into four broad classes commonly used in outcome research on humanistic psychotherapies (Elliott, 2001), each represented by at least two different samples: general clinical distress, specific symptoms, self-perception, and life functioning.

Between-Client Correlations. As indicated in Table 7, across the five data sets, 17 comparisons were carried out between data from the mean PQ and other outcome measures at the between-client level (level 2) with an overall weighted correlation of .51 (CI: .44 to .58), a large effect size that shows clear evidence of

convergence, but not so large as to indicate redundancy with standardized outcome measures. However, there is a moderate level of heterogeneity among these effects (Q = 36.48; df = 16; I^2 = 56.1%) indicating important variability among these effects. Seven of the 17 comparisons fell outside the confidence interval boundaries: effects from the USA outpatient sample appeared to be smaller than those for the other samples, and although we could not discern differences across measure type (see Table 7), correlations between the PQ and CORE-OM (the most frequently-used measure) were highly consistent and large (mean weighted r = .60; CI: .52 - .66; Q = 2.11; df = 3; NS; I^2 = 0.

Within-Client Correlations. Table 7 also presents the 15 within-client (level 1) multilevel correlations between PQ and standardized outcome measures, controlling for grouping of assessment data within clients: The overall weighted mean within-client correlation was .41 (95% CI: .25 - .55), a medium-to-large effect that also fell inside the hypothesized range for measure convergence. There was an even higher level of heterogeneity among these effects (Q = 96.05; df = 14; $p < .01$; I^2 = 85.4%). Five of these effects fell outside the confidence interval, and although measure type did not appear to make a difference in convergence, correlations in the USA outpatient sample again appeared to be somewhat smaller and were more likely to fall outside of the confidence interval for within-client convergence.

Discriminant Validity. Finally, although the data sets in this study were not designed to assess discriminant validity, we were able to tentatively examine this in the USA outpatient sample by looking past the Neuroticism scale of the NEO-FFI to its other 4 scales: Correlations between data from the PQ and these other NEO-FFI scales varied from -.10 (between-client r) and -.11 (within-client r) for Openness to Experience to -.30 (between client r) and -.27 (within client r) for Conscientiousness , with values for Extraversion (r between: -.25; r within: -24) and (r between: -.15; r within: -.17) for Agreeableness, all somewhat lower than the mean convergent correlations reported.

Sensitivity Analyses: Measuring Change with the PQ

Pre-post change. The PQ is used to measure change both pre-post and from session-to-session and so it is important to establish norms both (a) for effect sizes (Cohen's d), for power calculations in future studies, and (b) for calculating reliable change index (RCI) values (Jacobson & Truax, 1991), for estimating rates of client improvement and deterioration and identifying sessions with sudden gains or losses (Tang & Derubeis, 1999). (Sensitivity to change refers to differences in absolute level over time, whereas temporal consistency, reported earlier, involves relative stability of rank order standing over time; cf. Durbin & Klein, 2006.)

Table 8 contains pre-post scale-level outcome data from each of the five data sets, calculated conservatively by using paired-sample tests of all clients who received at least 3 sessions of therapy and PQ from final session of therapy. Standardized differences of the mean (Cohen's *d*) varied widely, from .82 (USA outpatient sample) to 1.69 (USA depression sample), with an overall value of 1.25 (n = 348; CI: .26 to 2.24). In addition, using pre-therapy sd and screening-to-session-1 test-retest correlations, we calculated reliable change index (RCI) values for $p < .05$ (Jacobson & Truax, 1991) for each sample; the overall weighted mean value was 1.67, somewhat higher than has been reported for measures of general psychological distress such as the CORE-OM (e.g., .59; Connell et al., 2007). Data indicating the proportion of clients showing this amount of pre-post improvement varied across samples, with general outpatient samples showing lower rates of improvement (lowest value was for

the USA outpatient sample) than the two samples focused on clients with specific presenting problems (highest value was USA depression); overall, slightly more than a third of clients showed reliable pre-post change (36%). Deterioration rates were uniformly low, averaging about 1% (range 0 to 4%).

Session-to-session change. Finally, we examined session-to-session weekly change on PQ scale-level scores at lag-1 (see Table 8). The overall weighted mean differences were small but positive (n = 5937; mean = .08; sd = .71), varying from .06 (USA outpatient) to .13 (USA depression). The overall value for the upper 95th percentile was 1.45, while the overall session-to-session RCI value was 1.40. The largest of these values was found for the Portuguese sample, while the smallest values occurred in the two Scottish samples.

Discussion

The overall purpose here was to establish a set of psychometric parameters for a simplified brief, individualized, client-generated outcome measure, the Personal Questionnaire. To this end, we analysed five data sets from three countries, including both English-language and Portuguese version. Reviewing the proposed hypotheses, we found that:

(1) Typical normative characteristics of PQ scores have been able to be established, including:
   (a) Number of items: The weighted mean number of rated items was found to be around 10, which matches the instrument construction guidelines.
   (b) Initial severity: Mean pre-therapy PQ scores averaged about 5 on the PQ's 7-point rating scale, indicating that the average client's average problem had bothered them "considerably" during the previous week.
   (c) Duration of problems: The mean duration of problems experienced at roughly the same level as this initial severity was reported as 3 to 5 years.

(2) PQ scores generally showed good internal consistency, varying from the .70's into the .80's, with pre-therapy between-client reliabilities a bit higher and more variable than within-client reliabilities.

(3) PQ scores over time were strongly consistent. Our best estimate of the temporal reliability PQ scores is .57, the between-client correlation between intake and session 1 (an average interval of about a month); this is the value recommended for calculating the reliable change interval (RCI).

(4) PQ scores showed strong correlations with standardized outcome measures at both between- and within-client levels, typically ranging between .3 and .6, including a range of other measures of clinical distress in different clinical populations, especially general distress, but also measures of self-perception and life functioning.

(5) PQ scores were able to detect client change session to session and over the course of therapy. Large pre-post standardized mean differences ranging from .8 to 1.7 were found, with the largest effects being for clients seen in focused, time-limited treatments for specific presenting problems (anxiety or depression). Clinical cut-off and reliable change threshold values have been able to be established:
   (a) Based on our results it now appears that a caseness threshold of 3.25 fits the data best and is a reasonable compromise between the 3.0 value used by Barkham et al. (1996) and the 3.5 used by Elliott et al. (2009).

(b) For calculating reliable change index (RCI) index values (Jacobson & Truax, 1991), we recommend a minimum of about 1.50 points (based on 1.67 points for pre-post change, and 1.4 points for week-to-week change at $p < .05$) to justify a claim of strong evidence that a client has shown significant change.

These results constitute a wide range of evidence supporting the psychometric quality of scores derived from the PQ. However, an unresolved question is whether these scores can be viewed as measures of a single, coherent personalized problem distress index. Our internal consistency analyses did indicate that in general alpha was adequate for this purpose. Still, there were clearly wide variations among clients in internal item consistency, including items inconsistent with such a general index, as well as multiple dimensions of individualized personal problem distress, which PCA undoubtedly over-estimated here. Clearly, there is a need for further research on the internal consistency and factor structure of PQ scores including using methods such as parallel analysis and MAP methods (e.g., O'Connor, 2000).

Several broad criticisms have been made of individualized outcome measures (Mintz & Kiesler, 1982; Ogles, Lambert, & Masters, 1996; Waskow & Parloff; 1975): Each client has a unique set of items, which may make it difficult to compare or average scores across clients. They are too specific and therefore may neglect other facets of change. They lack adequate psychometric data. Most tellingly, they are a time consuming method for assessing a general psychological distress. Some of these complaints are more relevant to the PQ than others. For instance, the simplified version of the PQ studied here was designed to be relatively brief and simple to administer and score, and the PQ appears to suffer less from problems of over-specificity compared to other individualized treatment measures.

Other criticisms lodged against the PQ require further examination. For example, the problem of item noncomparability across clients seems an obvious limitation, challenging the practice of calculating group mean scores in outcome studies or creating and using normative data to help interpret PQ scores. This noncomparability is documented not only by the highly diverse item content generated by different clients, but also by wide variations in patterns of inconsistent items and factor structures. Although these arguments do not address using the PQ to track the progress of individual clients, several responses to this general critique are possible: (a) If there are enough sessions, we recommend carrying out individualized factor analyses to group problems broadly by content and to make more specific comparisons over time. New methods for individualized within-client comparisons are now available (e.g., Metric-Frequency similarity methods; Sales & Wakker, 2009, available online in http://mfcalculator.celiasales.org/, and described by Sales, Wakker, Alves & Faísca, in press). (b) Conceptually, rank ordering means that each client's items overall have the meaning of "the most important problems that I want to work in psychotherapy," while particular items have the meaning of "the problem I initially ranked X in order of importance to work on." (c) Although the severity of PQ items ostensibly does differ between clients, other measures are also susceptible to the problems of lack of comparability across clients; that is, standard items may look the same but mean different things to different people, who also vary in their response sets.

In contrast, another criticism asserts that the PQ may simply be measuring a more general dimension, such as general psychological distress. Our analyses did indeed show relatively high correlations between data from the PQ and measures of global clinical distress (e.g., especially the CORE-OM). Our analyses suggest that

this overlap is probably greater in clinical trials focusing on particular client populations than in general clinical samples. On the other hand, it may also be that the PQ points to the specific client issues that give rise to client general psychological distress, in the same way that specific disease processes (e.g., viral load, injury) underlie signs of general immune system activation (e.g., inflammation).

Although the samples used here could be faulted on the basis of being mostly limited to humanistic-experiential psychotherapies and graduate-student therapists, in our view, the main methodological limitation of the data stems from its multilevel nature, involving extensive nonindependence of observations. We were able to document the nature of this nonindependence in our analyses of temporal structure; however, owing to the complexity of the data (especially the varying numbers of items across and within clients) we were only able to implement fully multilevel statistical analyses for the convergence analyses.

While these studies provide important information about the PQ, it would be helpful to utilize the PQ with other treatment approaches, to examine the impact of level of therapist training upon PQ scores, and to evaluate the temporal structure of PQ scores across more long-term therapies. Other potential areas for future research include comparing data from the PQ with other individualized treatment measures, more systematic testing of discriminant validity, examining PQ scores in a wider range of mental health settings and additional countries and languages, and examining how factors like gender, diagnosis and clinical setting affect PQ scores. Further research might also look at the factor structure of individual client PQs using more robust factor analytic methods to determine optimum number of factors (e.g., Parallel Analysis; Ledesma & Valero-Mora, 2007), or test whether moving average methods (e.g., averaging successive sets of three sessions) might improve temporal stability and therefore reduce the RCI interval, which is particularly relevant for case studies.

Most importantly, we did not examine PQ item content here. Although a simple content analysis system for classifying PQ item content exists (Barkham, Shapiro & Morrison, 1988), it was developed for use with a particular client population (depressed clients with work issues). Thus, a key issue is developing better methods for classifying PQ item content and using these to describe kinds of client presenting problem, based on a broader range of clients.

We conclude with a discussion of some of the broader advantages and uses of the PQ. Most generally, the PQ mixes and integrates idiographic and nomothetic approaches to assessment. It is individualized, while at the same time, as we have shown here, it can be understood and analyzed as a psychometric measure of psychological distress, so that different clients' scores can be compared normatively and combined in group studies (e.g., Barkham et al., 1989). This makes the PQ useful for group designs, including RCTs and practice-based research, as well as for the new generation of systematic case studies (McLeod, 2010).

In terms of clinical utility, the PQ has potential as a measure that can appeal to therapists from a wide range of theoretical orientations: Its specificity is consistent with cognitive-behavioral therapy; its personal, individualized nature fits well with psychodynamic and humanistic-experiential approaches; and it can easily accommodate the more systemic issues brought by clients in couple and family therapies. It is also highly consistent with collaborative assessment approaches developed by Fischer (1994) and Finn and Tonsager (1997). Beyond this, it appears that the PQ has a variety of uses in clinical practice. First, clients often find the process of constructing the PQ to be useful for clarifying their focus and goals for therapy. In addition to identifying a range of key problems that a client wants to work

on in psychotherapy, PQ items can be used as a basis for case formulation by looking at patterns of interrelated items. Finally, particular PQ items can be deconstructed as potential markers for specific kinds of therapeutic work within a given therapeutic perspective (e.g., interpersonal loss-oriented vs. self-critical/self-esteem issues in depression). In summary, we find the PQ to be a robust client generated outcome measure which has demonstrated sound psychometric properties as well as clinical utility.

References

Accurso, E.C., Hawley, K.M., & Garland, A. F. (2013). Psychometric properties of the Therapeutic Alliance Scale for caregivers and parents. *Psychological Assessment, 25*, 244-252. DOI: 10.1037/a0030551

Allport, G. (1937). *Personality: A psychological interpretation*. New York: Holt

Allport, G. (1960). *Personality and social encounter*. Boston: Beacon Press.

Alves, P., Sales, C., & Ashworth, M. (2013). Enhancing the patient involvement in outcomes: a study protocol of personalised outcome measurement in the treatment of substance misuse. *BMC Psychiatry, 13*, 337-349.

American Psychiatric Association. (1993). *Diagnostic and Statistical Manual - Fourth Edition*. Task Force on DSM-IV: Washington, D.C.

Ashworth, M., Shepherd, M., Christey, J., Matthews, V., Wright, K., Parmentier, H., Robinson, S., & Godfrey, E. (2004). A client-centred psychometric instrument: the development of PSYCHLOPS. *Counselling and Psychotherapy Research, 4*, 27-33.

Ashworth, M., Robinson, S., Evans, C., Shepherd, M., Conolly, A., & Rowlands, G. (2007). What does an idiographic measure (PSYCHLOPS) tell us about the spectrum of psychological issues and scores on a nomothetic measure (CORE-OM)? *Primary Care and Community Psychiatry, 12*, 7-16.

Barkham, M., Rees, A., Stiles, W.B., Shapiro, D.A., Hardy, G.E., & Reynolds, S. (1996). Dose-effect relations in time-limited psychotherapy for depression. *Journal of Consulting and Clinical Psychology, 64*, 927-935.

Barkham, M., Shapiro, D.A., Firth-Cozens, J. (1989). Personal questionnaire changes in prescriptive vs. exploratory psychotherapy. *British Journal of Clinical Psychology, 28*, 97-107.

Barkham, M., Shapiro, D. A., & Morrison, L. (1988). *Classification of psychological problems elicited by the Personal Questionnaire technique: A coding manual*. Sheffield, UK: University of Sheffield, MRC/ESRC Social and Applied Psychology Unit.

Barkham, M., Stiles, W.B., & Shapiro, D.A. (1993). The shape of change in psychotherapy: Longitudinal assessment of personal problems. *Journal of Consulting and Clinical Psychology, 61,* 667-677.

Bates, D., Maechler, M., Bolker. B.M., & Walker. S. (2014). *lme4: Linear mixed-effects models using Eigen and S4*. ArXiv e-print: http://arxiv.org/abs/1406.5823.

Borenstein, M., Hedges, L.V., Higgins, J.P.T., & Rothstein, H.R., (2009). *Introduction to Meta-Analysis*. New York: Wiley.

Carvalho, M. J., Faustino, I., Nascimento, A., & Sales, C. M. D. (2008). Understanding Pamina's recovery: An application of the hermeneutic single-case efficacy design. *Counseling and Psychotherapy Research, 8*, 166-173.

Connell, J., Barkham, M., Stiles, W.B., Twigg, E., Singleton, N., Evans, O., Miles, J.N.V. (2007). Distribution of CORE-OM scores in a general population, clinical cut-off points and comparison with the CIS-R. *British Journal of Psychiatry, 190*, 69-74. DOI: 10.1192/bjp.bp.105.017657

Connor, K.M., Davidson, J.R.T., Churchill, L.E., Sherwood, A., Foa, E., & Weisler, R.H. (2000). Psychometric properties of the Social Phobia Inventory (SPIN). *British Journal of Psychiatry, 176*, 379-386.

Costa, P.T., & McCrae, R.R. (1992). *The NEO PI-R*. Odessa, FL: PAR.

Derogatis, L. R. (1983). *SCL-90-R administration, scoring and procedures manual - II*. Towson, MD: Clinical Psychometric Research.

Durbin, E.C., & Klein, D. N. (2006). Ten-year stability of personality disorders among outpatients with mood disorders. *Journal of Abnormal Psychology, 115*, 75–84.

Egan, V., Miller, E., & McLellan, I. (1998). Does the personal questionnaire provide a more sensitive measure of cardiac surgery related-anxiety than a standard pencil-and-paper checklist? *Personality and Individual Differences, 24*, 451-583. DOI: 10.1016/S0191-8869(97)00192-X

Elliott, R. (2001). Research on the effectiveness of humanistic therapies: A meta-analysis. In D. Cain & J. Seeman (Eds.), *Humanistic psychotherapies: Handbook of research and practice* (pp. 57-81). Washington, D.C.: APA.

Elliott, R., Fox, C.M., Beltyukova, S.A., Stone, G.E., Gunderson, J., & Zhang, Xi. (2006). Deconstructing therapy outcome measurement with Rasch analysis: The SCL-90-R. *Psychological Assessment, 18*, 359-372.

Elliott, R., Mack, C., & Shapiro, D. (1999). *Simplified Personal Questionnaire Procedure*. (On-line). Available: http://www.experiential-researchers.org/instruments/elliott/pqprocedure.html

Elliott, R., Partyka, R., Alperin, R., Dobrenski, R., Wagner, J., Messer, S. B., Watson, J., & Castonguay, L. (2009). An adjudicated hermeneutic single-case efficacy design study of experiential therapy for panic / phobia. *Psychotherapy Research, 19*, 543-557

Elliott, R., Watson, J.C., Goldman, R.N., & Greenberg, L.S. (2004). *Learning emotion-focused therapy: The process-experiential approach to change*. Washington, DC: APA.

Evans, C., Connell, J., Barkham, M., Margison, F., Mellor-Clark, J., McGrath, G. & Audin, K. (2002). Towards a standardised brief outcome measure: Psychometric properties and utility of the CORE-OM. *British Journal of Psychiatry, 180*, 51-60.

Faur, A., & Elliott, R. (2007). *Self-Relationship Questionnaire*. Unpublished questionnaire, Counselling Unit, University of Strathclyde.

Finn, S.E., & Tonsager, M.E. (1997). Information-gathering and therapeutic models of assessment: complementary paradigms. *Psychological Assessment*, 9, 374–385. doi:10.1037/1040-3590.9.4.374

First, M.B., Spitzer, R.L, Gibbon, M., & Williams, J.B.W. (1997). *Structured Clinical Interview for DSM-IV Personality Disorders, (SCID-II)*. Washington, D.C.: American Psychiatric Press.

First, M.B., Spitzer, R.L., Gibbon, M., & Williams, J.B.W. (2002). *Structured Clinical Interview for DSM-IV-TR Axis I Disorders, Research Version, Non-patient Edition (SCID-I/NP).* New York: Biometrics Research, New York State Psychiatric Institute.

Fischer, C.T. (1994). *Individualizing psychological assessment*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Freire, E.S. (2007). *Development of a psychotherapy outcome measure based on Rogers' theory of therapy change*. Unpublished MSc thesis. Counselling Unit, University of Strathclyde.

Glass, G. V., Willson, V. L., & Gottman, J. M. (1975). *Design and analysis of time-series experiments*. Boulder, CO: Colorado Associated University Press.

Good, P. (2006) *Resampling Methods* (3rd Ed.). Boston: Birkhauser.

Gorsuch, R.L. (1997). Exploratory factor analysis: Its role in item analysis. *Journal of Personality Assessment, 68*, 532-560.

Grafanaki, S., & Mcleod, J. (1999). Narrative processes in the construction of helpful and hindering events in Experiential Psychotherapy. *Psychotherapy Research, 9*, 289-303.

Horowitz, L.M., Rosenberg, S.E., Baer, B.A., Ureño, G., Villaseñor, V.S. (1988). Inventory of interpersonal problems: psychometric properties and clinical applications. *Journal of Consulting and Clinical Psychology, 56*, 885-892.

Hunsley, J, & Mash, E.J. (2007). Evidence-based assessment. *Annual Review of Clinical Psychology, 3*, 29-51.

Jacobson, N.S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59*, 12-19.

Kiesler, D. J. (1966). Some myths of psychotherapy research and the search for a paradigm. *Psychological Bulletin, 65*, 110-136.

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*, 443-452.

Klein, M. J., & Elliott, R. (2006). Client accounts of personal change in process experiential psychotherapy: A methodologically pluralistic approach. *Psychotherapy Research, 16*, 91-105.

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9: validity of a brief depression severity measure. *Journal of General Internal Medicine, 16*, 606–613.

Lambert, M. J., & McRoberts, C. (1993). *Survey of outcome measures used in JCCP 1986-1991*. Poster presented at the Annual Meeting of the Western Psychological Association, Phoenix, Arizona.

Ledesma, R.D., & Valero-Mora, P. (2007). Determining the number of factors to retain in EFA: An easy-to-use computer program for carrying out Parallel Analysis. *Practical Assessment Research & Evaluation, 12*(2), 1–11.

Maling, M.S., Gurtman, M.B., & Howard, K.I. (1995). The response of interpersonal problems to varying doses of psychotherapy. *Psychotherapy Research, 5*, 63-75.

McLeod, J. (2010). *Case study research in counselling and psychotherapy*. London: Sage.

Macleod, R., Elliott, R., & Rodgers, B. (2012). Process-experiential / emotion-focused therapy for social anxiety: A hermeneutic single-case efficacy design study hermeneutic single-case efficacy design study. *Psychotherapy Research, 22*, 67-81.

McPherson, F.M., & LeGassicke, L.A. (1965). A single-patient self-controlled trial of Wy 3498. *British Journal of Psychiatry, 111*, 149-154.

Mellor-Clark, J. et al. (1998). *CORE System (Information Management) Handbook*. Leeds, UK: CORE System Group.

Messer, B., & Harter, S. (1986). *Manual for the adult self-perception profile*. Denver, CO: University of Denver.

Mintz, J., & Kiesler, D. (1982). Individualized measures of psychotherapy outcome. In P. Kendall& J.N. Butcher (Eds.), *Handbook of research methods in clinical psychology* (pp.491-534). New York: Wiley and Sons.

O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instrumentation, and Computers*, *32*, 396-402.

Ogles, B., Lambert, M., & Masters, K. (1996). *Assessing outcome in clinical practice.* Boston: Simon & Schuster.

Pascal, G.R., & Zax, M. (1956). Psychotherapeutics: Success or failure. *Journal of Consulting Psychology, 20*, 325-331.

Phillips, J.P.N. (1986). Shapiro personal questionnaire and generalized personal questionnaire technique: A repeated measures individualized outcome measurement. In L. Greenberg & W. Pinsof (Eds.). *The Psychotherapeutic Process: A Research Handbook* (pp. 557-589). New York: Guilford.

R Core Team. (2012). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.

Rickard, H.C. (1965). Tailored criteria of change in psychotherapy. *Journal of General Psychology, 72*, 63-68.

Robins, L.N., Helzer, J.E., Croughan, J., & Ratcliff, K.S. (1981). National Institute of Mental Health Diagnostic Interview Schedule: Its history, characteristics, and validity. *Archives of General Psychiatry, 38*, 381-389.

Rush, J., & Hofer, S.M. (2014). Differences in within- and between-person factor structure of positive and negative affect: Analysis of two intensive measurement studies using multilevel structural equation modeling. *Psychological Assessment, 26*, 462–473. DOI: 10.1037/a0035666

Sales, C. M. D., & Alves, P. C. G. (2012). Individualized patient-progress systems: Why we need to move towards a personalized evaluation of psychological treatments. *Canadian Psychology, 53*, 115-121.

Sales, C. M. D., & Alves, P. C. G. (2014). *Psychotherapy through the eyes of patients: A review of assessment tools*. Submitted for publication.

Sales, C.M.D., Alves, P.C.G., Evans, C., & Elliott, R. (2014). The Individualized Patient Progress System (IPPS): A decade of international collaborative networking. *Counselling and Psychotherapy Research, 14*, 181-191. DOI: 10.1080/14733145.2014.929417

Sales, C. M. D., Moleiro, C., Evans, C., & Alves, P. C. G. (2012). Versão Portuguesa do CORE-OM: Tradução, adaptação e estudo preliminar das suas propriedades psicométricas. *Revista de Psiquiatria Clínica, 39*, 54-59.

Sales, C. M. D., Gonçalves, S., Silva, I.F., Duarte, J., Sousa, D., Fernandes, E., Sousa, Z., & Elliott, R. (2007, March). *Portuguese adaptation of qualitative change process instruments.* Paper presented at the European Chapter Annual Meeting of the Society for Psychotherapy Research, Funchal, Portugal.

Sales, C. M. D., & Wakker, P. P. (2009). The metric-frequency measure of similarity for ill-structured data sets, with an application to family therapy. *The British Journal of Mathematical and Statistical Psychology, 62*, 663-82.

Sales, C. M. D., Wakker, P., Alves, P., & Faísca, L. (in press). MF Calculator: a web-based application for analyzing similarity. *Journal of Statistical Software.*

Shapiro, M.B. (1961). A method of measuring changes specific to the individual psychiatric patient. *British Journal of Medical Psychology, 34*, 151-155.

Shapiro, M.B. (1969). Short-term improvements in the symptoms of affective disorder. *British Journal of Social and Clinical Psychology, 8*, 187-188.

Snijders, T.A.B., & Bosker, R.J. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Los Angeles: Sage

Tang, T.Z., & DeRubeis, R.J., (1999). Sudden gains and critical sessions in cognitive-behavioral therapy for depression. *Journal of consulting and clinical psychology, 67*, 894-904. http://dx.doi.org/10.1037/0022-006X.67.6.894

Vieira, F., Torres, S., & Moita, G. (2011, September). *Psychological intervention in obesity: psychodramatic approach*. Paper presented at The 4th Regional Mediterranean Congress of the International Association for Group Psychotherapy and Group Processes, Porto, Portugal.

Waskow, I.E., & Parloff, M.B. (1975). *Psychotherapy change measures.* Rockville, MD: National Institute of Mental Health.

Wing, J.K., Cooper, J.E., & Sartorius, N. (1974). *Measurement and Classification of Psychiatric Symptoms; An instruction Manual for the PSE and Catego Program*. New York: Cambridge University Press.

Table 1

Overview of Study Sample Characteristics

	USA Depression	**USA Outpatient**	**Portugal Outpatient**	**Scotland Social Anxiety**	**Scotland Outpatient**	**Overall**
N of clients	48	64	72	64	207	455
% Female	77	58	86	55	67	
% European origin	92	90	94[a]	97	95	
Mean (SD) age	36.2 (11.1)	43.3 (13.3)	32.7 (10.1)	35.3 (10.4)	36.9 (11.9)	
N therapists	10	15	6	15	33	79
Type of therapy offered	EFT	EFT	Various	PCT or EFT	PCT	
Total N of sessions	559	934	872	1226	3516	7107
Sessions offered	20	40	Various	20	40	
Mean (SD) sessions offered	12.7 (6.1)	14.7 (15.2)	14.6 (12.6)	14.7 (7.3)	15.1 (12.9)	
Comparator instruments	SCL-90R	CORE-OM, NEO-FFI, GAF, Harter IIP-26,	CORE-OM, PHQ-9	CORE-OM, SPIN, SRQ, IIP-26, SI	CORE-OM, SI	19
Years data collected	1986-90	1998-2002	2011-12	2007-12	2007-13	

Note. EFT: Emotion-Focused Therapy; PCT: Person-Centered Therapy; see text for abbreviations of comparator instruments.

[a]Gave nationality as Portuguese

Table 2.
Overview of Analyses

Type of Analysis/Hypothesis	**Item vs. Scale Analysis**	**Within vs. Between Client level**	**Resampling**	**When in Treatment**	**Minimum n Observations per Client**	**Integration Model (weighting)**
(1) *Normative*: (a) n items; (b) initial severity; (c) duration [see Table 3]	Item: 1a, 1c; Scale: 1b	Within: 1a, 1c; Between: 1b	No	All sessions: 1a; Pre-therapy: (usually intake) 1b; Intake: 1c	1 session or intake: 1a, 1b; Intake: 1c	Weighted mean (sample size)
(2) *Internal structure:* (a) Internal consistency, (b) inconsistent items, (c) dimensionality [Tables 4 & 5]	Item: all	Between: 2a; Within: all	2 – 13 items: 2a, 2b; 2 – 12 items: 2c	Pre-therapy: 2a (Between); Weekly & intake: all (Within)	≥20 clients: 2a (Between); ≥10 sessions: all (Within)	Alphas: Random effects (inverse variance);. Other statistics: Weighted mean (sample size)
(3) *Temporal structure:* (a) pre-therapy test-retest correlations, (b) pre-post correlations, (c) session-to-session autocorrelations, (d) variance accounted for by clients, (e) time series analyses [Table 6]	Scale: all	Within & Between (pooled): all	No	Intake & session 1: 3a Pre- & Post: 3b; All sessions: 3c, 3d, 3e; Up to session 20: 3e	Intake & session 1: 3a; ≥3 sessions: 3b; 2 successive sessions: 3c, 3e[a]; 1 session: 3d	Random effects (inverse variance)
(4) *Construct validity*: correlations with standardized outcome measures of psychological distress	Scale	Within & Between (multilevel)	No	All assessments (screening, pre, mid, post,	1 assessment	Random effects (inverse variance)

[Table 7]				follow-up)		
(5) *Sensitivity to change:* (a) pre-post change; (b) session-to-session change [Table 8]	Scale: all	Within & Between (pooled): all	No	Pre- & Post: 5a; All sessions: 5b	3 sessions: all	Effect sizes: Random effects (inverse variance); Other statistics: Weighted mean (sample size)

[a]Up to 20 sessions used for nonstationarity analyses (3e)

Table 3

Descriptive Data for Personal Questionnaires Across Samples

	USA Depres-sion	**USA Outpatient**	**Portugal Outpatient**	**Scotland Social Anxiety**	**Scotland Outpatient**	**Overall**
Number of Items (averaged across sessions):						
n	559	934	872	1226	3516	7107
m	9.4	9.9	5.1	9.9	10.4	9.5
sd	2.7	3.5	2.2	2.9	2.8	2.8
range	7 to 12	4 to 23	1 to 13	4 to 26	4 to 16	1 to 26
Pretreatment Severity (averaged across clients):						
n	45	63	67	64	188	427
m	5.25	4.94	4.89	5.07	5.07	5.04
sd	.68	1.13	1.15	.84	.84	.93
range	3.75 to 6.64	1.38 to 7	1 to 7	3 to 7	2.6 to 7	
Prior Duration of Problems (averaged across clients):						
n	--	21	72	63	196	352
m	--	4.75	3.78	6.00	5.12	4.98
sd	--	1.24	1.90	.88	1.25	1.36
range	--	2.1 to 6.43	1 to 7	3.67 to 7	1.86 to 7	

Note. Overall figures use weighted means and pooled sds. Duration ratings: 1: less than 1 month; 2: 1 - 5 months; 3: 6 - 11 months; 4: 1 - 2 years; 5: 3 - 5 years; 6: 6 - 10 years; 7: more than 10 years. Duration ratings not used with USA Depression sample.

Table 4

Between-client Internal Reliability of Personal Questionnaire Scores at Pre-therapy

	USA Depression	**USA Outpatient**	**Portugal Outpatient**	**Scotland Social Anxiety**	**Scotland Outpatient**	**Overall**
m alpha	.71	.87	.84	.76	.77	.80
sd alpha	.10	.03	.01	.10	.08	*.03*[a]
Range of client n	38 – 43	22 - 56	23 – 71	29-64	28-188	43-155
Range of items with clients n ≥ 20 sessions	2 - 10	2-11	2 – 6	2 - 10	2 - 13	2 - 13

Note. Alphas calculated repeatedly for k = 2 – 13 items, with varying numbers of clients (depending on number of items analyzed) as long as n ≥ 20 clients; overall alpha and standard error calculated using weighted meta-analytic random effects model.

[a]Standard error of the cross-sample weighted mean

Table 5
Within-Client Internal Structure of Personal Questionnaire Scores Across Sessions: Reliability and Number of Factors

	USA Depression	**USA Outpatient**	**Portugal Outpatient**	**Scotland Social Anxiety**	**Scotland Outpatient**	**Overall**
N clients with 10+ sessions	30	30	29	42	105	236
N of items: m (sd)	10.7 (1.1)	9.0 (2.5)	5.0 (1.8)	9.3 (2.0)	9.9 (2.3)	9.2 (2.1)
Alpha at max items: m (sd)	.79 (.24)	.74 (.30)	.75 (.19)	.74 (.26)	.79 (.25)	.77 *(.03[a])*
% Clients with alpha ≥ .7 at max items	77%	70%	66%	86%	79%	77%
N incon-sistent items: m (sd)	2.0 (2.6)	1.7 (1.8)	1.0 (1.4)	1.7 (2.5)	1.7 (2.5)	1.7 (2.3)
% clients with ≤ 2 inconsistent items	70%	70%	86%	76%	73%	74%
N factors at max items: m (sd)	2.7 (.9)	2.6 (.9)	1.8 (.8)	2.2 (1.1)	2.5 (1.2)	2.4 (1.1)
% 1-factor solutions	10%	10%	41%	34%	21%	23%

Note. Alpha calculated separately for each client, at maximum number of items for which there was data from at least 10 sessions (n items = 2 – 12) for each client; overall alpha and standard error calculated using weighted meta-analytic random effects model. Inconsistent items defined as corrected item-total < .3. Number of factors estimated by principal components analyses (eigenvalue = 1 criterion) calculated for maximum number of items for clients for which there were 10+ data points. Overall figures use weighted means and pooled sds.
[a]Standard error of the mean

Table 6
Temporal Structure and Pooled Time Series Analyses of Personal Questionnaire Scores Across sessions

	USA Depression	**USA Outpatient**	**Portugal Outpatient**	**Scotland Social Anxiety**	**Scotland Outpatient**	**Overall**
(a) Intake w Session 1 correlation (N)	--	.39** (55)	.73** (55)	.57** (59)	.54** (164)	.57** (333) *.06*[a]
Mean days intake-session 1	--	34	13.1	47.8	35.4	33.8
(b) Pre-post correlation (N)	.46** (45)	.47** (55)	.35* (53)	.41** (52)	.38** (140)	.41** (345) *.05*[a]
(c) Lag-1 auto-correlation overall (N)	.79** (512)	.83** (844)	.73** (614)	.86** (1133)	.85** (3309)	.82** (6412) *.02*[a]
(d) Correlation with session number (N)	-.45** (536)	-.09* (655)	-.16** (675)	-.44** (933)	-.21** (2831)	-.28** (5630) *.07*[a]
(e) Lag-1 auto-correlation of differenced scores (N)	-.42** (450)	-.38** (762)	-.40** (551)	-.35** (1051)	-.37** (3116)	-.37** (5930) *.01*[a]
(f) Eta-squared for variance attributable to clients (F-values)	.52 (12.42**)	.68 (28.42**)	.52 (13.45**)	.55 (21.23**)	.61 (25.02**)	.58 *.03*[a]

* $p < .05$; **$p < .01$ for pooled client analyses

Note. (a), (b) & (c) provide three different estimates of temporal stability (test-retest reliability). PQ not rated at intake in USA Depression sample. (b) Pre-post correlations calculated for cases with ≥ 3 sessions between session 1 and last available score. (d) Assesses degree of nonstationarity overall. (Sessions < 21 used for (d).) (e) indicates presence of autoregressive process in differenced scores. (f) Eta-squared analyses assess overall level of statistical nonindependence within cases (associated F value in parentheses). Overall correlations and standard error calculated using weighted meta-analytic random effects model.

[a]Standard error of the mean

Table 7

Between- and Within-Clients Correlations between PQ Scores and Other Measures of Psychological Distress or Functioning

Type of Measure	Measure	Sample (n)	Between Clients Correlation		Within Clients Correlation	
			r w PQ	N (clients)	r w PQ	N (assessments)
General Clinical Distress	CORE-OM	USA Outpatient	.53**	27	.26**	62
		Portugal Outpatient	.54**	71	--	--
		Scotland Social Anxiety	.54**	57	.48**	207
		Scotland Outpatient	.64**	214	.58**	536
	SCL90-R	USA Depression	.65**	27	.69**	111
		USA Outpatient	.45**	59	.33**	159
	NEO Neuroticism	USA Outpatient	.36**	54	.40**	151
	GAF Therapist Rating	USA Outpatient	-.17	25	-.25	34
	Weighted Mean (CI)		*.49** (.34 - .61)*	*8 effects*	*.41 (.25 - .55)*	*7 effects*
Specific Symptoms	PHQ-9	Portugal Outpatient	.44**	58	--	--
	Social Phobia Inventory	Scotland Social Anxiety	.73**	64	.68**	232
	Weighted Mean (CI)		*.61** (.25 - .82)*	*2 effects*	*.68** (.60 - .74)*	*1 effect*
Self Perception	Harter Self-Concept[a]	USA Outpatient	-.50**	55	-.36**	153
	Self-Relationship Scale: Self Affiliation[a]	Scotland Social Anxiety	-.47**	53	-.54**	172
	Self-Relationship Scale: Self Attack	Scotland Social Anxiety	.51**	53	.49**	171
	Weighted Mean (CI)		*.49** (.36 - .60)*	*3 effects*	*.47** (.36 - .56)*	*3 effects*
Functioning	Inventory of	USA Outpatient	.30*	59	.18**	164

	Interpersonal Problems					
		Scotland Social Anxiety	.64**	52	.49**	168
	Strathclyde Inventory[a]	Scotland Social Anxiety	-.52**	57	-.54**	205
	Strathclyde Inventory[a]	Scotland Outpatient	-.53**	212	-.52**	516
	Weighted Mean (CI)		*.51** (.38 - .61)*	*4 effects*	*.44** (.30 - .57)*	*4 effects*
Overall Weighted Mean (CI)			*.51** (.44 - .58)*	*17 effects*	*.41** (.25 - .55)*	*15 effects*

* p < .05; **p < .01.

[a]Instrument scored in positive (non-distressed) direction (sign reversed for meta-analysis).

Note. CI: 95% confidence interval. Between-client correlations used mean client scores, weighted by number of assessments (df = n clients – 2). Within-client correlations used multilevel correlation controlling for nonindependence within clients (R mle4 package; df = n assessments – k clients - 1). Overall correlations and their confidence intervals calculated using weighted meta-analytic random effects model, with correlations with measures of positive functioning reversed.

Table 8

Sensitivity Analyses: Measuring Change with the PQ

Sample:	**USA Depression**	**USA Out-patient**	**Portugal Out-patient**	**Scotland Social Anxiety**	**Scotland Out-patient**	**Overall**
Pre-Post Change						
N clients	45	55	56	52	140	348
Pre-test m (sd)	5.22 (.71)	5.10 (.96)	4.50 (1.30)	5.50 (.82)	4.98 (.84)	5.03 (.93)
Post-test m (sd)	3.51 (1.24)	4.17 (1.28)	3.34 (1.30)	3.78 (1.32)	3.75 (1.41)	3.72 (1.34)
ES (signif)	1.69**	.82**	.89**	1.57**	1.06**	1.25*
RCI ($p < .05$)	1.45	2.08	1.87	1.49	1.58	1.67
% reliable improvement	48.9%	25.5%	30.4%	42.3%	36.4%	36.2%
% reliable deterioration	0%	3.6%	1.8%	0%	.7%	1.1%
Session-to-Session Change						
N clients	512	844	723	1133	2725	5937
Mean diff between sessions at lag 1 (sd)	.13** (.81)	.06* (.79)	.12** (.86)	.07** (.63)	.07** (.64)	.08** (.71)
Upper 95%-ile RCI ($p < .05$)	1.72 1.56	1.61 1.54	1.80 1.75	1.30 1.22	1.32 1.30	1.45 1.40

*$p < .05$; **$p < .01$, paired samples t-tests.

Note. All cases with ≤3 sessions used; ES: standardized mean difference. RCI: reliable change interval. For pre-post change, RCI estimates used pre-therapy sd and screening-to-session-1 correlation (estimated from the other 4 samples for USA Depression sample). % reliable improvement/deterioration used overall RCI value of 1.67. For weekly change, RCI estimates used total sample sd and lag 1 autcorrelations. Overall ES calculated meta-analytically using random effects model.