Psychological Assessment
2016, Vol. 28, No. 3, 263–278

1040-3590/16/$12.00 http://dx.doi.org/10.1037/pas0000174

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

Robert Elliott
University of Strathclyde

John Wagner
Dialectical Behaviour Therapy Centre of Vancouver, Vancouver, British Columbia, Canada

Célia M. D. Sales
Universidade de Évora and ISCTE – Instituto Universitário de Lisboa

Brian Rodgers
University of Queensland

Paula Alves
ISCTE – Instituto Universitário de Lisboa

Maria J. Café
Portuguese Association of Family and Community Therapy, Lisbon, Portugal

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 5 data sets from 3 countries. Overall pretherapy mean internal consistency (alpha) across clients was .80, and within-client alphas averaged .77; clients typically had 1 or 2 items that did not vary with the other items. Analyses of temporal structure indicated high levels of between-clients variance (58%), moderate pretherapy test–retest correlation ($r = .57$), and high session-to-session Lag-1 autocorrelation (.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 cutoff of 3.25 and a reliable change index 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

Every client has a unique clinical condition, with a set of problems and presentations specific to his or her 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 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 article, 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 that "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 2 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,

This article was published Online First June 15, 2015.

Robert Elliott, School of Psychological Science and Health, University of Strathclyde; John Wagner, Dialectical Behaviour Therapy Centre of Vancouver, Vancouver, British Columbia, Canada; Célia M. D. Sales, Centro de Investigação em Educação e Psicologia, Universidade de Évora, and Centro de Investigação e Intervenção Social, ISCTE – Instituto Universitário de Lisboa; Brian Rodgers, School of Nursing, Midwifery and Social Work, University of Queensland; Paula Alves, Centro de Investigação e Intervenção Social, ISCTE – Instituto Universitário de Lisboa, Portugal; Maria 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.

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 Célia M. D. Sales and SFRH/BD/87308/2012 to Paula Alves). We thank Micaela Jiménez and Carlos Jiménez for statistical assistance.

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

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; MacLeod, Elliott, & Rodgers, 2012). This review also identified three main CGOMs in use today: Goal Attainment Scaling (GAS; Kiresuk & Sherman, 1968), Psychological Outcome Profiles (PSYCHLOPS; Ashworth et al., 2004), and the simplified version of the PQ (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 using 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, although the GAS has been widely used, it lacks psychometric research. (However, Ashworth et al., 2007, provided some limited psychometric analyses for PSYCHLOPS.)

The original PQ, 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 & Le Gassicke, 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, or researcher) helps the client through a process of developing a list of approximately 10 problem statements, describing in his or her own words what he or she wants to work on in treatment; the client then rates these problems on a seven-point scale. The process of constructing the list of PQ problem statements generally takes about 30 min 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-case 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 (Sales, Alves, Evans, & Elliott, 2014).

When it comes to the clinical utility of the PQ, a small study by Sales et al. (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 postsession discussions (75%). Among the advantages of the PQ, therapists reported usefulness for session-to-session outcome monitoring (38%), enhancement of 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%). 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 has been very little 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 alphas. 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 et al. (1989) reported evidence for reasonable convergence with data generated from two other symptom measures, the revised Symptom Checklist–90 (SCL-90-R; 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 ($r = .37$). Barkham et al. (1996) did look at the current version of the PQ but reported only pretherapy 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 ranging from 2.47 to 1.89.

Thus, in spite of such measures' 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 PQ data. In this article, we present and integrate data from five different data sets from three countries (the United States, Scotland, and Portugal), including both general outpatient and specialized client populations (depressed, socially anxious). Specifically, we examine the following sets of psychometric propositions or hypotheses to generate a network of evidence regarding the use of PQ scores in psychotherapy outcome assessment: (a) Normative: Typical quantitative characteristics of PQ scores can be established, including number of items, initial severity, and duration of problems. (b) Internal structure: PQ scores will show substantial levels of internal consistency ($\alpha \geq .70$), will have relatively few inconsistent items (≤ 2), and will be generally ($>50\%$) unidimensional (which would support the use of a single index of weekly client problem distress). (c) Temporal structure: PQ scores over time will be strongly consistent, showing large pretherapy 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. (d) Construct validity: PQ scores will show moderate-to-strong correlations (in the .40–.60 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 (>.70); in addition, for the small number of discriminant-validity associations assessed, we expected less than strong relationships (<.40). (e) 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 cutoff 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 to derive overall estimates of these parameters. The five samples come from five different psychotherapy outcome studies carried out between 1986 and 2013 by various combinations of the authors. 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; comparator instruments; and years of data collection—is provided in Table 1. Because the PQ was used across all five samples, it is 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 (generally 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. He or she then helped the client separate complex statements, clarified ambiguous statements, and encouraged the client to discard redundant statements to arrive at a list of approximately 10 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 seven-point anchored scale (1 = *not at all*, 2 = *very little*, 3 = *little*, 4 = *moderately*, 5 = *considerably*, 6 = *very considerably*, 7 = *maximum possible*). Finally, the client was asked to rate problem duration, also on a seven-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 U.S. depression data set.) Afterward, the client's PQ was typed up, leaving space for him or her to note any additional difficulties that he or she might subsequently experience. On subsequent administrations, clients rated severity (for the past week) only.

Table 1
Overview of Study Sample Characteristics

Variable	United States depression	United States outpatient	Portugal outpatient	Scotland social anxiety	Scotland outpatient	Total
Clients (*n*)	48	64	72	64	207	455
Female (%)	77	58	86	55	67	
European origin (%)	92	90	94[a]	97	95	
Age (years): *M* (*SD*)	36.2 (11.1)	43.3 (13.3)	32.7 (10.1)	35.3 (10.4)	36.9 (11.9)	
Therapists (*n*)	10	15	6	15	33	79
Type of therapy offered	EFT	EFT	Various	PCT or EFT	PCT	
Total sessions (*n*)	559	934	872	1,226	3,516	7,107
Sessions offered (*n*)	20	40	Various	20	40	
Sessions delivered (*n*): *M* (*SD*)	12.7 (6.1)	14.7 (15.2)	14.6 (12.6)	14.7 (7.3)	15.1 (12.9)	
Comparator instrument(s)	SCL-90-R	CORE–OM; NEO-FFI; GAF; Harter; IIP-26	CORE–OM; PHQ-9	CORE–OM; SPIN; SRQ; IIP-26; SI	CORE–OM; SI	
Years data collected	1986–1990	1998–2002	2011–2012	2007–2012	2007–2013	

Note. EFT = emotion-focused therapy; PCT = person-centered therapy; SCL-90-R = revised Symptom Checklist–90; CORE–OM = Clinical Outcome Routine Evaluation—Outcome Measure; NEO-FFI = NEO Five-Factor Inventory; GAF = Global Assessment of Functioning; Harter = Harter Global Self-Worth Scale; IIP-26 = Inventory of Interpersonal Problems–26; PHQ-9 = Patient Health Questionnaire–9; SPIN = Social Phobia Inventory; SRQ = Self-Relationship Questionnaire; SI = Strathclyde Inventory.
[a] Gave nationality as Portuguese.

U.S. 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, & Ratcliff, 1981) administered by trained research staff, all of the clients either met *Diagnostic and Statistical Manual of Mental Disorders* (*DSM*; 3rd ed. [American Psychiatric Association, 1980]) 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, and 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 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.)

U.S. 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 percent 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* (American Psychiatric Association, 1994) Axis I and Axis II disorders. A small number of clients were excluded, however, because they were actively suicidal, were 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 Structured Clinical Interview for *DSM–IV* Personality Disorders [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 EFT.

Measures. Clients completed the following (higher scores indicate greater distress or dysfunction unless otherwise stated): (a) The Harter Global Self-Worth Scale is a six-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 indicated greater levels of global self-worth. (b) The Inventory of Interpersonal Problems–26 (IIP-26; Horowitz, Rosenberg, Baer, Ureño, & 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) The 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 four other subscales (for which higher scores indicate more extraversion, openness to experience, agreeableness, or conscientiousness). (For this sample, internal alphas were as follows: Neuroticism: .85; Extraversion: .84; Openness to Experience: .77; Agreeableness: .68; and Conscientiousness: .87.) (d) The Clinical Outcome Routine Evaluation—Outcome Measure (CORE–OM; Evans et al., 2002; CORE System Group, 1998) is a standardized 34-item self-report measure of psychological distress using a five-point anchored frequency scale ranging from 0 (*not at all*) to 5 (*most or all of the time*), with a 1-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, 1994) ratings were completed by therapists at beginning and end of therapy. (f) Clients also completed the SCL-90-R, which was also used in U.S. depression sample (internal alpha for this sample was .97).

Portugal Outpatient Sample

Participants. A convenience sample following a practice-based research approach was constructed by inviting three free psychotherapy services of varying lengths (the University Counseling Service of the University of Madeira; the Department of Psychiatry of São João Hospital, Porto, Portugal; and the Psychotherapy Service of the Higher Institute for Applied Psychology, Lisbon, Portugal) and two independent-practice psychodrama group therapists (who ran groups of varying lengths) to join an online 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 (all female) participated, and 72 patients were recruited. Most of the clients (71%) had applied for individual therapy, and the rest were beginning psychodrama (29%). The majority of clients 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 at the pretreatment 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. On consent, the PQ interview took place to create the client-generated list of items, together with the CORE–OM and the Patient Health Questionnaire–9 (PHQ-9; Kroenke, Spitzer, & Williams, 2001). 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 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) The Portuguese version of the CORE–OM (Sales, Moleiro, Evans, & Alves, 2012), used also used in the U.S outpatient sample, was administered. (Internal consistency for this sample was .93.) (b) The PHQ-9 is a nine-item self-report measure evaluating the *DSM–IV* criteria for depression on a four-point anchored scale ranging 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 was .87.)

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; 64 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-I; 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 therapists in counseling (2), counseling psychology (2), or clinical psychology (1).

Procedure. Clients were offered up to 20 free sessions of either person-centered therapy or EFT 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 posttherapy, 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 was used, as in the U.S and Portugal outpatient samples (internal alpha for this sample was .95). (b) The Social Phobia Inventory (SPIN; Connor et al., 2000) is an 11-item problem-specific measure of social anxiety symptoms supported by evidence for good reliability and validity (internal alpha for this sample was .93). (c) The IIP-26 is an interpersonal problem distress measure, also used in the U.S. outpatient sample (internal consistency for this sample was .90). (d) The Strathclyde Inventory (SI; Freire, 2007) is an experimental person-centered outcome measure assessing a single 31-item dimension of congruence/fluidity versus incongruence/structure-boundness (internal alpha: .93), scored in the direction of higher client functioning. (e) The Self-Relationship Questionnaire (SRQ; Faur & Elliott, 2007) is an experimental instrument 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).

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 (PCE) 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 0–44 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 six male; 15 were diploma-level student counselors, 16 were counseling psychology doctoral students, and two were doctoral-level practitioners in either counseling or counseling psychology.

Measures. This sample used a subset of the measures used in the Scotland social anxiety sample: In addition to the PQ, these were the CORE–OM (internal alpha for this sample was .95) and the SI (internal alpha for this sample was .96).

Analysis Approach

Weekly PQ scores are multilevel 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. Accurso, Hawley, & Garland, 2013; Elliott et al., 2006), it is essential to assess the psychometric characteristics of scores at within-client as well as between-clients levels and to address issues of nonindependence. The multilevel nature of the data thus required a complex analytic strategy: Where possible (e.g., convergent validity analyses), we used multilevel analyses; however, in other cases (interitem 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 are noted. A summary overview of the five sets of analyses is provided in Table 2.

Results

Descriptive Analyses

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 7,107 sessions (see Table 3). Across samples, the weighted mean number of items rated was around 10 ($M = 9.5$, $SD = 2.8$), with only the Portuguese sample varying substantially ($M = 5.1$, $SD = 2.2$), probably because of 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; $M = 5.04$, $SD = 0.93$), which corresponds to "considerably" distressed. This value can be used for interpreting client initial PQ scores, for example, by using it to establish a caseness cutoff or threshold value according to Jacobson Criterion A (Jacobson & Truax, 1991), defined as pretreatment clinical population mean minus 2 standard deviations (i.e., <5% probability of belonging to a normative clinical population). Applying this criterion to our data yielded a clinical cutoff 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 partway through collection of the U.S. 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—that is, 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 sample reported the shortest duration of problems (roughly 1–2 years; see Table 3).

Internal Reliability Analyses

To assess internal consistency of item scores we (a) looked separately at both between-clients and within-client levels and (b) set minimum numbers of observations (either clients or sessions within clients) 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 level. To examine internal consistency at the between-clients level for each sample, pretherapy PQ scores were analyzed across clients repeatedly for 2–13 items (as long as n was at least 20 clients, a somewhat arbitrary value selected to increase stability of estimates). In other words, we used a resampling strategy (Good, 2006) in which we started with two 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 ($SE = .03$), with the lowest value for the U.S. depression sample (.71) and the highest value for the U.S. outpatient sample (.87).

Within-client level. To obtain reasonably stable estimates of interitem internal reliability at the within-client level, PQ data from each client were separately calculated using the maximum number of items for which there were 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 the most consistent item ratings (86%). Unsurprisingly, given natural clinical complexity and how PQ items are constructed for nonredundancy, clients typically had one or two items that were not internally consistent with the rest of the PQ items, defined by corrected item-total correlations of less than .30 (overall mean number of inconsistent items = 1.7, $SD = 2.3$). The Portuguese sample had the smallest number of inconsistent items ($M = 1.0$, $SD = 1.4$); the U.S. depression sample had the largest number ($M = 2.0$, $SD = 2.6$).

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; 2–12 items; casewise deletion of data); for each client, we selected the solution with the maximum number of items such

Table 2
Overview of Analyses

Type of analysis/hypothesis	Item vs. scale analysis	Within-client vs. between-clients level	Resampling	When in treatment	Minimum observations per client (*n*)	Integration model (weighting)
1. Normative: (a) items (*n*), (b) initial severity, and (c) duration (see Table 3)	Item:1a,1c Scale: 1b	Within: 1a, 1c Between: 1b	No	All sessions: 1a Pretherapy: (usually intake) 1b Intake: 1c	One session or intake: 1a, 1b Intake: 1c	Weighted mean (sample size)
2. Internal structure: (a) internal consistency, (b) inconsistent items, and (c) dimensionality (see Tables 4 and 5)	Item: All	Between: 2a Within: all	2–13 items: 2a and 2b 2–12 items: 2c	Pretherapy: 2a (between) Weekly and 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) pretherapy test–retest correlations, (b) pre–post correlations, (c) session-to-session autocorrelations, (d) variance accounted for by clients, and (e) time-series analyses (see Table 6)	Scale: All	Within and between (pooled): All	No	Intake and Session 1: 3a Pre–post: 3b All sessions: 3c, 3d, and 3e Up to Session 20: 3e	Intake and Session 1: 3a ≥3 sessions: 3b Two successive sessions: 3c and 3e[a] One session: 3d	Random effects (inverse variance)
4. Construct validity: Correlations with standardized outcome measures of psychological distress (see Table 7)	Scale	Within and between (multilevel)	No	All assessments (screening, pre, mid, post, follow-up)	One assessment	Random effects (inverse variance)
5. Sensitivity to change: (a) pre–post change and (b) session-to-session change (see Table 8)	Scale: all	Within and between (pooled): All	No	Pre–post: 5a All sessions: 5b	Three sessions: all	Effect sizes: Random effects (inverse variance) Other statistics: weighted mean (sample size)

[a] Up to 20 sessions used for nonstationarity analyses.

Table 3
Descriptive Data for Personal Questionnaires Across Samples

Variable	United States depression	United States outpatient	Portugal outpatient	Scotland social anxiety	Scotland outpatient	Overall
			Number of items (averaged across sessions)			
n	559	934	872	1,226	3,516	7,107
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	7–12	4–23	1–13	4–26	4–16	1–26
			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	0.68	1.13	1.15	0.84	0.84	0.93
Range	3.75–6.64	1.38–7.00	1–7	3–7	2.60–7.00	
			Prior duration of problems (averaged across clients)[a]			
n		21	72	63	196	352
M		4.75	3.78	6.00	5.12	4.98
SD		1.24	1.90	0.88	1.25	1.36
Range		2.10–6.43	1–7	3.67–7.00	1.86–7.00	

Note. Overall figures use weighted means and pooled standard deviations. Duration ratings were not used with United States depression sample.
[a] 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.

that there were 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, which 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 (*M* = 1.8, *SD* = 1.4), likely because of 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 U.S. 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 [RCI] 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 U.S. outpatient sample to .73 for the Portuguese sample, with an overall value of .57 (n = 353; 95% confidence interval [CI] [.43, .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, .49]).

Table 4
Between-Client Internal Reliabilities of Personal Questionnaire Scores at Pretherapy

Variable	United States depression	United States outpatient	Portugal outpatient	Scotland social anxiety	Scotland outpatient	Overall
Alpha (*M*)	.71	.87	.84	.76	.77	.80
Alpha (*SD*)	.10	.03	.01	.10	.08	.03[a]
Range of clients (*n*)	38–43	22–56	23–71	29–64	28–188	43–155
Range of items with clients ≥20 sessions (*n*)	2–10	2–11	2–6	2–10	2–13	2–13

Note. Alphas were calculated repeatedly for k = 2–13 items, with varying numbers of clients (depending on number of items analyzed) as long as $n \geq 20$ clients; overall alpha and standard error were calculated using a 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: Reliabilities and Numbers of Factors

Variable	United States depression	United States outpatient	Portugal outpatient	Scotland social anxiety	Scotland outpatient	Overall
Clients with 10+ sessions (*n*)	30	30	29	42	105	236
Items (*n*): *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 maximum items: *M* (*SD*)	.79 (.24)	.74 (.30)	.75 (.19)	.74 (.26)	.79 (.25)	.77 (.03[a])
Clients with alpha ≥.70 at maximum items (%)	77	70	66	86	79	77
Inconsistent items (*n*): *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
Factors at maximum items (*n*): *M* (*SD*)	2.7 (0.9)	2.6 (0.9)	1.8 (0.8)	2.2 (1.1)	2.5 (1.2)	2.4 (1.1)
One-factor solutions (%)	10	10	41	34	21	23

Note. Alphas were calculated separately for each client at the maximum number of items for which there were data from at least 10 sessions (*n* items = 2–12) for each client; overall alpha and standard error were calculated using a weighted meta-analytic random-effects model. Inconsistent items were defined by corrected item-total correlations <.30. Number of factors was estimated by principal components analyses (eigenvalue = 1 criterion), calculated for the maximum number of items for clients for which there were 10+ data points. Overall figures use weighted means and pooled standard deviations.
[a] Standard error of the mean.

Time-series parameters. Weekly administration of the PQ creates time-series data with potentially complex mathematical structures of nonindependence; these need to be understood 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-clients variance in each data set (see Table 6, Row f). Overall variance attributable to clients (eta-squared [random-effects model]) was .58 (n = 7,107; 95% CI [.52, .63]); the only sample whose eta-squared value fell outside this confidence interval was the U.S. outpatient sample, which was higher (η^2 = .68). Next, we looked at session-to-session (Lag-1) within-client autocorrelations, in which successive scores (here pooled but with breaks between clients) were correlated. These values (see Table 6, Row c), which also pick up both between-clients variance and secular trend, were also quite large and generally consistent across samples, with an overall weighted correlation of .82 (n = 6,412; 95% CI [.78, .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 autoregressive integrated moving average modeling (Glass, Willson, & 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

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

Variable	United States depression	United States outpatient	Portugal outpatient	Scotland social anxiety	Scotland outpatient	Overall
(a) Intake with Session 1 correlation		.39**	.73**	.57**	.54**	.57**
n		55	55	59	164	333
SE						.06
Mean days intake–Session 1		34	13.1	47.8	35.4	33.8
(b) Pre–post correlation	.46**	.47**	.35*	.41**	.38**	.41**
n	45	55	53	52	140	345
SE						.05
(c) Lag-1 autocorrelation overall	.79**	.83**	.73**	.86**	.85**	.82**
n	512	844	614	1,133	3,309	6,412
SE						.02
(d) Correlation with session number	−.45**	−.09*	−.16**	−.44**	−.21**	−.28**
n	536	655	675	933	2,831	5,630
SE						.07
(e) Lag-1 auto-correlation of differenced scores	−.42**	−.38**	−.40**	−.35**	−.37**	−.37**
n	450	762	551	1,051	3,116	5,930
SE						.01
(f) Eta-squared for variance attributable to clients	.52	.68	.52	.55	.61	.58
F	12.42**	28.42**	13.45**	21.23**	25.02**	
SE						.03

Note. Rows a, b, and c provide three different estimates of temporal stability (test–retest reliability). The Personal Questionnaire was not rated at intake in United States depression sample. Row b presents pre–post correlations calculated for cases with ≥3 sessions between Session 1 and the last available score. Row d assesses degree of nonstationarity overall (sessions <21 were used). Row e indicates the presence of autoregressive process in differenced scores. In Row f, eta-squared analyses assess overall level of statistical nonindependence within cases. Overall correlations and standard errors were calculated using a weighted meta-analytic random-effects model.
For pooled client analyses: * $p < .05$. ** $p < .01$.

d). There was a fair amount of variability across samples, with an overall weighted mean r of −.28 (n = 5,630; 95% CI [−.40, −.15]), a medium effect size, indicating that PQ scores were in general moderately nonstationary. However, the U.S. outpatient sample showed less consistent improvement associated with session number (r = −.09), whereas both the 20-session time-limited samples focused on particular client presenting problems (U.S. depression and Scotland social anxiety clients) showed higher levels of nonstationarity (rs = −.45 and −.44, respectively). This pointed to the value of working with the difference between successive PQ scores 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-clients 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 = 5,630; 95% CI [−.40, −.35]), with only the U.S. 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

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 used 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 midtherapy 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 to pick up change over time, but it results in multilevel data sets that need to be deconstructed (e.g., Rush & Hofer, 2014). Accordingly, we first analyzed between-clients 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, Mächler, 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-clients and within-client correlations. 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-clients 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-clients level (Level 2), with an overall weighted correlation of .51 (95% CI [.44, .58]), a large effect size showing clear evidence of convergence but not so large as to indicate redundancy with standardized outcome measures. However, there was a moderate level of heterogeneity among these effects (Q = 36.48, df = 16, I^2 = 56.1%), indicating important variability among them. Seven of the 17 comparisons fell outside the confidence interval boundaries: Effects from the U.S. 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 the CORE–OM (the most frequently used measure) were highly consistent and large (mean weighted r = .60, 95% 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 U.S. 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 U.S. outpatient sample by looking past the Neuroticism scale of the NEO-FFI to its other four scales: Correlations between data from the PQ and these other NEO-FFI scales varied from −.10 (between-clients r) and −.11 (within-client r) for Openness to Experience to −.30 (between-clients r) and −.27 (within-client r) for Conscientiousness, with values for Extraversion (between-clients r = −.25, within-client r = −.24) and Agreeableness (between-clients r = −.15, within-client r = −.17) 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 ds), for power calculations in future studies, and (b) for calculating 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.)

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 three sessions of therapy and the PQ from the final session of therapy, are presented in Table 8. Standardized differences of the mean (Cohen's ds) varied widely, from 0.82 (U.S. outpatient sample) to 1.69 (U.S. depression sample), with an overall value of 1.25 (n = 348; 95% CI [0.26, 2.24]). In addition, using pretherapy standard deviation and screening-to-Session-1 test–retest correlations, we calculated RCI values for

Table 7
Between-Clients and Within-Client Correlations Between Personal Questionnaire Scores and Other Measures of Psychological Distress or Functioning

			Between-clients correlation		Within-client correlation	
Type of measure	Measure	Sample	*r*	Clients (*n*)	*r*	Assessments (*n*)
General clinical distress	CORE–OM	United States outpatient	.53**	27	.26**	62
		Portugal outpatient	.54**	71		
		Scotland social anxiety	.54**	57	.48**	207
		Scotland outpatient	.64**	214	.58**	536
	SCL-90-R	United States depression	.65**	27	.69**	111
		United States outpatient	.45**	59	.33**	159
	NEO-FFI Neuroticism	United States outpatient	.36**	54	.40**	151
	GAF therapist rating	United States outpatient	−.17	25	−.25	34
	Weighted mean [95% CI]		.49** [.34, .61]	Eight effects	.41 [.25, .55]	Seven effects
Specific symptoms	PHQ-9	Portugal outpatient	.44**	58		
	Social Phobia Inventory	Scotland social anxiety	.73**	64	.68**	232
	Weighted mean [95% CI]		.61** [.25, .82]	Two effects	.68** [.60, .74]	One effect
Self-perception	Harter self-concept[a]	United States 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 [95% CI]		.49** [.36, .60]	Three effects	.47** [.36, .56]	Three effects
Functioning	Inventory of Interpersonal Problems	United States outpatient	.30*	59	.18**	164
		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 [95% CI]		.51** [.38, .61]	Four effects	.44** [.30, .57]	Four effects
Overall weighted mean [95% CI]			.51** [.44, .58]	17 effects	.41** [.25, .55]	15 effects

Note. Between-clients 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 (mle4 package in R; $df = n$ assessments − k clients − 1). Overall correlations and their confidence intervals (CIs) were calculated using a weighted meta-analytic random-effects model, with correlations with measures of positive functioning reversed. CORE–OM = Clinical Outcome Routine Evaluation—Outcome Measure (CORE–OM); SCL-90-R = revised Symptom Checklist–90; NEO-FFI = NEO Five-Factor Inventory; GAF = Global Assessment of Functioning.
[a] Instrument scored in the positive (nondistressed) direction (sign reversed for meta-analysis).
* $p < .05$. ** $p < .01$.

Table 8
Sensitivity Analyses: Measuring Change With the Personal Questionnaire

Variable	United States depression	United States outpatient	Portugal outpatient	Scotland social anxiety	Scotland outpatient	Overall
			Pre–post change			
Clients (*n*)	45	55	56	52	140	348
Pretest: *M* (*SD*)	5.22 (0.71)	5.10 (0.96)	4.50 (1.30)	5.50 (0.82)	4.98 (0.84)	5.03 (0.93)
Posttest: *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 (significance)	1.69**	0.82**	0.89**	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	0.7	1.1
			Session-to-session change			
Sessions (*n*)	512	844	723	1133	2725	5937
Difference between sessions at Lag 1: *M* (*SD*)	.13** (.81)	.06* (.79)	.12** (.86)	.07** (.63)	.07** (.64)	.08** (.71)
Upper 95th percentile	1.72	1.61	1.80	1.30	1.32	1.45
RCI ($p < .05$)	1.56	1.54	1.75	1.22	1.30	1.40

Note. All cases with ≤3 sessions were used. For pre–post change, reliable change index (RCI) estimates used pretherapy standard deviation and screening-to-Session-1 correlation (estimated from the other for samples for the United States depression sample). Percentages of reliable improvement and deterioration used an overall RCI value of 1.67. For weekly change, RCI estimates used total sample standard deviations and Lag-1 autcorrelations. Overall ES (standardize mean difference) was calculated meta-analytically using a random-effects model.
Paired-samples *t* tests: * $p < .05$. ** $p < .01$.

$p < .05$ (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., 0.59; 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 (the lowest value was for the U.S. outpatient sample) than the two samples focused on clients with specific presenting problems (the highest value was for the U.S. depression sample); overall, slightly more than one third of clients showed reliable pre–post change (36%). Deterioration rates were uniformly low, averaging about 1% (range: 0%–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 ($n = 5{,}937$; $M = 0.08$, $SD = 0.71$), varying from 0.06 (U.S. outpatient sample) to 0.13 (U.S. depression sample). The overall value for the upper 95th percentile was 1.45, whereas the overall session-to-session RCI value was 1.40. The largest of these values was found for the Portuguese sample, whereas 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 CGOM, the PQ. To this end, we analyzed five data sets from three countries, including both English-language and Portuguese versions. Reviewing the proposed hypotheses, we found the following:

1. Typical normative characteristics of PQ scores have been able to be established, including the following: (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 pretherapy PQ scores averaged about 5 on the PQ's seven-point rating scale, indicating that the average client's average problem had bothered him or her "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 3–5 years.

2. PQ scores generally showed good internal consistency, varying from the .70s into the .80s, with pretherapy between-clients 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-clients correlation between intake and Session 1 (an average interval of about a month); this is the value recommended for calculating the RCI.

4. PQ scores showed strong correlations with standardized outcome measures at both between-clients and within-client levels, typically ranging between .30 and .60, 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 0.8 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 cutoff and reliable change threshold values have been able to be established: (a) On the basis of 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 RCI values (Jacobson & Truax, 1991), we recommend a minimum of about 1.50 points (based on 1.67 points for pre–post change and 1.40 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 overestimated here. Clearly, there is a need for further research on the internal consistency and factor structure of PQ scores, including research using methods such as parallel analysis and minimum average partial 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 are 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 overspecificity compared with 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 at 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 on in psychotherapy," whereas 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. However, 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 delivered by 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.

Although these studies provide important information about the PQ, it would be helpful to use the PQ with other treatment approaches, to examine the impact of level of therapist training on 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 data from 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 important, 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 problems on the basis of 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, it can be understood and analyzed as a psychometric measure of psychological distress, so 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 randomized clinical trials 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 couples 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 focuses 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 orientation vs. self-critical/self-esteem issues in depression). In summary, we find the PQ to be a robust CGOM that 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. http://dx.doi.org/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. C. G., Sales, C. M. D., & 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. http://dx.doi.org/10.1186/1471-244X-13-337

American Psychiatric Association. (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed.). Washington, DC: Author.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

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 & Community Psychiatry, 12,* 7–16.

Ashworth, M., Shepherd, M., Christey, J., Matthews, V., Wright, K., Parmentier, H., . . . Godfrey, E. (2004). A client-centred psychometric instrument: The development of PSYCHLOPS. *Counselling & Psychotherapy Research, 4,* 27–31. http://dx.doi.org/10.1080/14733140412331383913

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. http://dx.doi.org/10.1037/0022-006X.64.5.927

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. http://dx.doi.org/10.1111/j.2044-8260.1989.tb00820.x

Barkham, M., Shapiro, D. A., & Morrison, L. (1988). *Classification of psychological problems elicited by the Personal Questionnaire technique: A coding manual*. Sheffield, England: 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. http://dx.doi.org/10.1037/0022-006X.61.4.667

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. (2014). *Fitting linear mixed-effects models using lme4*. Retrieved from 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. http://dx.doi.org/10.1002/9780470743386

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. *Counselling & Psychotherapy Research, 8,* 166–173. http://dx.doi.org/10.1080/14733140802211002

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. http://dx.doi.org/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). New self-rating scale. *British Journal of Psychiatry, 176,* 379–386. http://dx.doi.org/10.1192/bjp.176.4.379

CORE System Group. (1998). *CORE Systems (information management) handbook*. Rugby, UK: Author.

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

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

Durbin, C. E., & Klein, D. N. (2006). Ten-year stability of personality disorders among outpatients with mood disorders. *Journal of Abnormal Psychology, 115,* 75–84. http://dx.doi.org/10.1037/0021-843X.115.1.75

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,* 465–473. http://dx.doi.org/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, DC: American Psychological Association.

Elliott, R., Fox, C. M., Beltyukova, S. A., Stone, G. E., Gunderson, J., & Zhang, Xi. (2006). Deconstructing therapy outcome measurement with Rasch analysis: The Symptom Checklist-90-Revised. *Psychological Assessment, 18,* 359–372. http://dx.doi.org/10.1037/1040-3590.18.4.359

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

Elliott, R., Partyka, R., Alperin, R., Dobrenski, R., Wagner, J., Messer, S. B., . . . Castonguay, L. G. (2009). An adjudicated hermeneutic single-case efficacy design study of experiential therapy for panic/phobia. *Psychotherapy Research, 19,* 543–557. http://dx.doi.org/10.1080/10503300902905947

Elliott, R., Watson, J. C., Goldman, R. N., & Greenberg, L. S. (2004). *Learning emotion-focused therapy: The process-experiential approach to change*. Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/10725-000

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. *British Journal of Psychiatry, 180,* 51–60. http://dx.doi.org/10.1192/bjp.180.1.51

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

Finn, S. E., & Tonsager, M. E. (1997). Information-gathering and therapeutic models of assessment: Complementary paradigms. *Psychological Assessment, 9,* 374–385. http://dx.doi.org/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, DC: 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: Erlbaum.

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

Glass, G. V., Willson, V. L., & Gottman, J. M. (1975). *Design and analysis of time-series experiments*. Boulder: 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. http://dx.doi.org/10.1207/s15327752jpa6803_5

Grafanaki, S., & McLeod, J. (1999). Narrative processes in the construction of helpful and hindering events in experiential psychotherapy. *Psychotherapy Research, 9,* 289–303. http://dx.doi.org/10.1080/10503309912331332771

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. http://dx.doi.org/10.1037/0022-006X.56.6.885

Hunsley, J., & Mash, E. J. (2007). Evidence-based assessment. *Annual Review of Clinical Psychology, 3,* 29–51. http://dx.doi.org/10.1146/annurev.clinpsy.3.022806.091419

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. http://dx.doi.org/10.1037/0022-006X.59.1.12

Kiesler, D. J. (1966). Some myths of psychotherapy research and the search for a paradigm. *Psychological Bulletin, 65,* 110–136. http://dx.doi.org/10.1037/h0022911

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,* 443–453. http://dx.doi.org/10.1007/BF01530764

Klein, M. J., & Elliott, R. (2006). Client accounts of personal change in process-experiential psychotherapy: A methodologically pluralistic approach. *Psychotherapy Research, 16,* 91–105. http://dx.doi.org/10.1080/10503300500090993

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9. *Journal of General Internal Medicine, 16,* 606–613. http://dx.doi.org/10.1046/j.1525-1497.2001.016009606.x

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

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,* 1–11.

MacLeod, R., Elliott, R., & Rodgers, B. (2012). Process-experiential/emotion-focused therapy for social anxiety: A hermeneutic single-case efficacy design study. *Psychotherapy Research, 22,* 67–81. http://dx.doi.org/10.1080/10503307.2011.626805

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. http://dx.doi.org/10.1080/10503309512331331146

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

McPherson, F. M., & Le Gassicke, J. (1965). A single-patient, self-controlled and self-recorded trial of Wy 3498. *British Journal of Psychiatry, 111,* 149–154. http://dx.doi.org/10.1192/bjp.111.471.149

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.

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, Instruments, & Computers, 32,* 396–402. http://dx.doi.org/10.3758/BF03200807

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

Pascal, G. R., & Zax, M. (1956). Psychotherapeutics: Success or failure. *Journal of Consulting Psychology, 20,* 325–331. http://dx.doi.org/10.1037/h0040582

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 Press.

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. http://dx.doi.org/10.1001/archpsyc.1981.01780290015001

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. http://dx.doi.org/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/Psychologie canadienne, 53,* 115–121. http://dx.doi.org/10.1037/a0028053

Sales, C. M. D., & Alves, P. C. G. (2014). *Psychotherapy through the eyes of patients: A review of assessment tools*. Manuscript 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 & Psychotherapy Research, 14,* 181–191. http://dx.doi.org/10.1080/14733145.2014.929417

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

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 [Portuguese translation of CORE-OM: Translation, adaptation and preliminary study of its psychometric properties]. *Revista de Psiquiatria Clínica, 39,* 54–59. http://dx.doi.org/10.1590/S0101-60832012000200003

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. *British Journal of Mathematical and Statistical Psychology, 62,* 663–682. http://dx.doi.org/10.1348/000711008X376070

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 psychological changes specific to the individual psychiatric patient. *British Journal of Medical Psychology, 34,* 151–155. http://dx.doi.org/10.1111/j.2044-8341.1961.tb00940.x

Shapiro, M. B. (1969). Short-term improvements in the symptoms of affective disorder. *British Journal of Social & Clinical Psychology, 8,* 187–188. http://dx.doi.org/10.1111/j.2044-8260.1969.tb00606.x

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

Received April 3, 2014
Revision received April 23, 2015
Accepted April 27, 2015 ■