Title: Validation of the French Versions of Two Brief, Clinician-friendly Outcome Monitoring Tools: The ORS and SRS.

Short title: *VALIDATION OF THE FRENCH VERSION OF TWO BRIEF*

Christophe Cazauvieilh, Ph.D.*[1], Kamel Gana, Ph.D.[2], Scott D. Miller, Ph.D. [3], Bruno Quintard, Ph.D.[4]

[1] University of Bordeaux & IRPPSY, France
[2] University of Bordeaux, France
[3] International Center for Clinical Excellence, Chicago, Illinois, United Stated of America
[4] University of Bordeaux, France

*Corresponding author information : Christophe Cazauvieilh, 5 rue du Chai des Farines, 33000 Bordeaux – France, e-mail : christophe.cazauvieilh@u-bordeaux.fr

Abstract:

While Routine Outcome Monitoring (ROM) of change and therapeutic alliance is proving to be a promising way addressing the issues of drop-out and client deterioration in psychotherapy, at present ROM is infrequently employed in French-speaking contexts. This study aimed at testing psychometric properties of the French versions of two popular and widely-used ROM tools from the Partners for Change Outcome Management System (PCOMS): The Outcome Rating Scale (ORS) and Session Rating Scale (SRS).

The project implied a multiple study survey with an online data collection from clinical and non-clinical samples and investigated reliability, convergent validity, and factorial structure of the two scales, with in addition preliminary estimates regarding the clinical cut-off and reliable change index for the ORS.

The ORS and SRS have good psychometric properties regarding their brevity. Reliability, normative data, a reliable change index and scale thresholds are reported for the French versions.

The clinical use of PCOMS in French-speaking contexts could be encouraged bringing significant concrete elements of answer to the current stakes regarding access, regulation and delivering of psychotherapy in French-speaking areas.

Keywords: Partners for Change Outcome Management System (PCOMS); Outcome Rating Scale (ORS); Session Rating Scale (SRS); Psychometric qualities, Alliance, Feedback

Introduction

Evidence supporting the effectiveness of psychotherapy for health is well-established (Lambert & Ogles, 2004). Complimentary data indicate that not all clients benefit from psychotherapy (with approximately 50% end service before achieving a reliable change), that a significant percentage drops out ([~25%] Swift & Greenberg, 2012), and that a consistent percentage is worse off following care (5% to 10% of adult clients participating in clinical trials; Lambert et al., 2004).

Routine Outcome Monitoring System (ROM) has emerged as a method for addressing the issues of retention, deterioration, and lack of improvement (Lambert, 2015). The process involves providing normative feedback to clinicians generated from session-by-session monitoring of scores on standardized outcome and alliance measures. In a systematic review of different systems, Miller & Schuckard (2016) concluded that ROM can reduce drop-out as much as 50%, lessen deterioration rates by 33%, and double the rate of reliable and clinically significant change.

While multiple specific psychotherapeutic feedback systems are used worldwide by agencies and clinicians, and the efficiency of this system has been reported (Cazauvieilh, 2014, 2015), to date no comprehensive ultra-brief and evidence-based ROM system, including alliance and outcomes measures had been validated in French and ROM is infrequently used in French-speaking contexts.

One group of measures in wide use around the world is the Partners for Change Outcome Management System (PCOMS; Miller et al., 2005), including two scales: The Outcome Rating Scale (ORS) and the Session Rating Scale (SRS). Both tools have been tested rigorously, proven to provide reliable and valid scores, and are listed on the National Registry of Evidence-

Based Practice and Programs (EBP-NREPP) in the USA[1]. Additionally, both of these measures are brief in nature, requiring little time to administer, score, and interpret. This is critical as any measures or combination of measures that take more than five minutes to administer and to interpret are much less likely to be used in routine practice (Brown, Dreis, & Nace, 1999; Miller et al., 2003; Seidel et al., 2016). Finally, and most important, the feedback provided by these two specific tools has been tested in a number of randomized controlled trials and been shown to improve effectiveness and retention as well as decrease the risk of deterioration.

The present study aimed at testing French versions of the scales ORS and SRS and specifically, investigating their reliability, convergent validity, and factorial structure. In addition, using data from both clinical and non-clinical sample, the study seeks to establish preliminary estimates regarding the clinical cut-off and reliable change index for the ORS.

The first scale to investigate, the Outcome Rating Scale (ORS; Miller et al., 2005) is an ultra-brief, client-completed questionnaire designed to assess various dimensions well-being (e.g., individual, relationship, social, and global). The questionnaire includes four items displayed in a visual analog format. Each is scored on a scale from 0 to 10, with higher scores indicating higher levels of well-being. The scale is administered at the beginning of each session to determine baseline levels as well as track improvement in well-being.

The second scale, the Session Rating Scale (SRS 3.0; Miller et al., 2002) is also a-brief, client-completed instrument. It was designed to assess dimensions of the therapeutic relationship using Bordin's definition of the therapeutic alliance (1979, Miller et al., 2003). Similar to the ORS, the SRS contains four items displayed in a visual analog format assessing: (a) the quality of the bond/relation between client and therapist, (b) agreement on the goals for treatment, (c) the agreement about the tasks of therapy, and (d) the global quality of the session. Each item is scored from 0 to 10, with higher scores indicating a stronger therapeutic alliance.

[1] (http://www.nrepp.samhsa.gov/ViewIntervention.aspx?id=249).

Material and methods

Translation procedure of the ORS and SRS

Methodology followed the standards (International Test Commission 2014; Valleyrand , 1989). In 2013, the first author led a technical committee working to develop a unified French version of the measures. Then a pretest for clarity was conducted by the first author with 20 people in a French speaking University (heterogeneous sample). The French scales were well received. Worth noting that ORS item d ("Overall") was rated at a satisficing but lower level.

Participants and Procedure

Clinical samples

A first clinical sample composed of 28 clients entering the first session of psychotherapy was recruited through four Psychologists (3 women, 1 man), who showed interest in participating to our study. Initially seven therapists showed an interest to participate in our study. Both psychologists and their clients were fully informed about the protocol and signed respectively an electronic informed consent for clients, and a paper-pencil version for a therapist.

Data have been securely surveyed online at each session on practitioners' device (computer or tablet) by collecting non-identifying data and attributing a code to follow each participant. Neither the investigator had access to client identity, nor the psychologist accessed to client records, rendering a non-feedback condition (NFC). In this group, therapist caseload regarding the study ranged from 1 to 16 clients, client age ranged from 20 to 45 years (M = 30.43, SD = 6). 71.4% participants were in a close relationship. Regarding education and

occupation level, 96.4% of this group were at the French Baccalaureate level or above, and 85.7% were reporting currently working.

Three out of four therapists provided additional data about themselves and their practice: therapist age (M = 52.33, SD = 5.86), their clinical experience in years (M = 8.33, SD = 2.08), the treatment approach used (Behavioral and Cognitive therapy was the most frequent treatment), and information about their 24 respective clients; 16.7% of them were labeled "mandated to care" (not coming into therapy directly from themselves), 79.2% patients came with a mood/and or an anxiety disorder with a sub-proportion of 4.2% patients presenting concurrently anxiety and substance abuse issues, 20.8% came for other reasons including health/somatic issues and personality disorders. 37.5% of clients were receiving psychotropic drugs. Available client's session ranged from 1 to 9. Therapists were advised to stop recording if the client was returning to care after duration superior to 8 weeks since his/her last session (12.50 % of cases). Patients dropped out of their treatment in 37.50% of cases; at the end of the data collection remaining patients were either still in treatment (25%) or finished therapy (25%).

A second clinical sample (n = 28) composed of patients from the first author outpatient site practice, who began their first episode of treatment using the PCOMS French unified version, qualifying the research for a multiple study. These clients were informed that the therapist routinely used feedback scales to monitor outcome progress, the therapeutic relationship and adapt the treatment, rendering a feedback condition (FC) and signed an informed consent allowing the use of their data. ORS and SRS were administered at each session. Participant's problems were self-reported and assessed by the therapist (not categorized as in the NFC group). Mode treatment was Behavioral and Cognitive therapy. Client age ranged from 18 to 71 years (M = 34.61, SD = 12.89), 71.4% reported being in couples. 53.6% were currently at work. 57.2% were coming for mood, anxiety or emotional problems, 21.4% for

relational issues, 7.1% for professional issues, and 14.3% for other reasons (e.g., stress, addiction, health issues). 42.9% of people were identified to receive a psychotropic treatment at a given moment. Available client's session ranged from 1 to 15 (only the first episode of treatment was considered, with a delay between last session-next session not exceeding 8 weeks or 61 days).

Inclusion and Exclusion Criteria

To participate, it was required to be at or over 18 years old, and to attend a first session of therapy with the psychologist. People presenting issues incompatible with the administration of the measures were excluded (e.g., psychosis, severe difficulties with the French language and impairments complicating the self-administration of the measures).

Community Sample

To compare scores from a clinical population with a non-clinical community sample, a student sample was constituted by aggregating student samples from two French speaking universities. Volunteer filled out an online informed consent at first occasion (test) and filled out the survey at a second occasion (retest) on a personal computer or a tablet. A total of 154 students filled out the questionnaire on test. Among them, twenty-four reported currently receiving psychotherapeutic and/or psychotropic treatment (Clinical Community Sample, CCS; for subsidiary analysis) and 130 did not (Non-Clinical Community Sample, NCCS). There were 83.8% of women and 16.2% of men in NCCS, 91.7% of men and 8.3% of men in CCS. Students age ranged from 18 to 59 (M = 27.82, SD = 11.20) in NCCS and from 18 to 51 (M = 30.92, SD = 11.58) in CCS. In NCCS 45.38% of people were in couples (living together, married, civil partners. In CCS, 50% were in couples. As college students, the totality of NCCS and CCS participants was above French baccalaureate education level and 53.1% of NCCS students reported not having a job versus 37.5% in CCS group.

The first author used the LimeSurvey platform to render administration mostly similar in computerized presentation to paper-pencil, to secure and to standardize web administration. In these three samples, only complete data for each session record were investigated (if the participant filled out all the questions related to the session). Non-feedback samples data (clinical and non-clinical) came from a doctoral project and from routine clinical practice.

Compliance with ethical standards. Research Involving Human Participants

This research complied with French and international ethical standards regarding data collection in human science and was declared to the French CNIL registry. Informed consent (written or electronically) was required from all participants (therapists and clients) before the data collection.

Measures

In addition to the administration of PCOMS and contextual variables, our protocol contained the following measures:

The Outcome Questionnaire-22 (OQ-22), a self-administered instrument composed of 22 items, designed to assess the degree of distress experimented by clients during therapy and therefore the degree of change related to the treatment (OQ-HAI-22, Lambert & Burlingame, 1996b). The measure is composed of questions regarding distress in three clinical dimensions (symptom distress, interpersonal relationship, and social roles). Each item is in a five-point Likert scale response format, with score ranging from 0 to 4. The OQ-22 total score ranges from 0 to 88. Higher score indicates higher distress. OQ-22 is a reduced version of the "gold-standard measure" of psychotherapeutic outcomes Outcome Questionnaire- 45 (OQ-45.2, Lambert & Burlingame, 1996a). We used the French version of the OQ-22 validated by Flynn et al. (2003) to examine convergent validity with the ORS.

The Working Alliance Inventory-Short form (WAI-S), a 12 items self-administered scale, designed to assess the quality of the relationship between therapist and his patient, according to Bordin's model (1979, Tracey & Kokotovic,1989). The WAI-S is composed of questions assessing the three clinical dimensions of the working alliance (bond, agreement on goals, agreement on tasks). Each Likert style answer is on a seven-point format. Total scores are ranging from 0 to 84, where higher scores are the manifestation of a better therapeutic alliance. We used the French version of the WAI-S validated by Corbière et al. (2006) to explore convergent validity with the SRS, after replacing the term "case manager " with "therapist", following author recommendations, and slightly reformulated items 5 and 12 to better fit the French formulation.

The Kessler-6 Scale (K6), an ultra-brief self-administered questionnaire designated to assess general clinical distress (Kessler and al., 2003; National population Health survey, 1999), with six items and Likert style answer on a five-point format. The total scores are ranging from 0 to 24, with higher scores indicating higher distress. The K6 scale has been used worldwide as a clinical distress screening tool, and in French publications (Nguyen and al., 2012). We used the French authorized version edition of this scale to examine convergent validity with the ORS (A French validity study of the K6 has been led by Arnaud et al. (2010)). However, after consulting the author of the original scale (Pr R Kessler), we modified the instructions to consider only the 7 last days.

The Big Five Inventory (BFI), a self-administered scale designed to assess personality traits according to the big five models with the dimensions of Extraversion, Agreability, Conscientiousness, Negative affectivity, and Openessness (John, Donahue & Kettle, 1991). The measure is composed of 45 items with Likert style answer on a five-point format. Higher scores on a dimension are an indication of more marked presence. We used the French version of the BFI validated by Plaisant et al. (2010) to assess criterion validity of the ORS.

Statistical analyses

Because limited sample size, a complementary approach to estimate effects was selected (parametric, non-parametric and/or robust procedures). Differences between clinical samples (NFC and FC) have been assessed using Mann-Whitney test (2-sided) at session 1 (with bootstrapp for confidence interval (CI; BCa)). We also investigated differences in scores between Feedback and Non-Feedback Condition.

Content and convergent validity were examined by using Spearman rho tests for correlations with bootstrap CI at session 1 with NFC scores (on ORS and SRS) and on the first occasion (test) with the NCSS scores (ORS). Scales reliability has been computed at session 1 by using Cronbach alpha in a group composed of NFC and FC clients (for ORS and SRS) and in the NCCS at occasion1 (test). Construct validity has been investigated by comparing differences in ORS scores between clinical and non-clinical samples on occasion 1 (session 1 or test) and between sessions 1 and 2 or first session (intake) and last session (sensitivity to change by the test-retest with Wilcoxon rank test); and for SRS scores by examining correlation with subsequent ORS scores. Confirmatory Factorial analyses (CFA) were performed to assess factorial structure of ORS and SRS. Preliminary norms have been documented (ORS, SRS, occasion 1) and change index (ORS; cut-off, reliable change index or RCI; SRS, threshold) have been examined by using Jacobson and Truax's formulas (1991). Finally, we drew ROC curves with ORS scores at session 1(clinical and non-clinical). Effect sizes have been analyzed according to Cohen's recommendations (2013).

Results

Outcome Rating Scale (ORS)

Descriptive Statistics

Table 1 presents means and standard deviations of the ORS scores in each group. The ORS total scores in clinical groups (NFC and FC) are lower than those reported by Miller et al. (2003; *M* =19.6, *SD* =8.7) but comparable to those reported in other European validation studies by Janse et al. (2014; *M* = 17.0, *SD* = 7.2) and Moggia et al. (2017; 2018; *M* =18.18, *SD* = 9.91).The scores in the community sample (NCCS) are lower than those reported by Miller et al. (2003; *M* =28 *SD* =6.8), and lower than those reported by Bringhurst et al. (2007; *M* = 29.9, *SD* = 7.5) and by Janse et al. (2014; *M* = 29.6, *SD*=6.0).Women intake ORS scores in the Non-Feedback Clinical group (NFC; *Mean Rank* = 13.52) did not significantly differ from men scores (NFC; *Mean Rank* = 17.43, *U* = 94, *p* = .30, *d* = .21). Women intake scores in the Feedback Clinical group (FC; *Mean Rank* = 13.06) did not significantly differ from scores from men (*Mean Rank* = 16.42, *U* =73, *p* =.30, *d* = 0.20). Scores at session 1 in the Non-Clinical Community Sample for Women (NCCS; *Mean Rank* = 67.19) did not differ significantly from scores from men (*Mean Rank* = 56.71, *U* = 960, *p* = .24 *d* = .10).

--INSERT TABLE 1 HERE--

Scale and item analysis

Factorial Structure

Smith and al. (2010) indicated one factor structure of the ORS by an Exploratory Factorial Analysis (ORS subscales loading ranged respectively .69, .62, .76, .98). We performed a Confirmatory Factorial Analysis (CFA), including the whole student sample at occasion 1 (test; *n* =154). A one-factor model was estimated in Mplus version 7.4 (Muthén &

Muthén, 2010), using a robust procedure (i.e., MLR). The results revealed an improper solution because of a negative residual variance in the theta matrix (Heywood case) regarding ORS item d ("Overall"), which was not significant (p> .05). The same issue has been encountered with estimating in the TCS and the NFC sample, but not with the Feedback NFC sample.

As the negative residual variance for ORS item d was small and nonsignificant (*-.05, P> .05*), we followed Muthen recommendation[2] and constrained the variance on ORS item d to 0. The model fairly fits well the data ($\chi^2(2) = 3.89$, $p = .27$, CFI = 1, TLI = 0.98; RMSEA = 0.04, 90% CI [.00, .15] ; respective items factor loadings : 88, .54, .60, 1).

Reliability

As there were no significant differences in ORS total scores at session 1 between clients in the NFC group (n =28, *Mean Rank* = 29.48) and clients in the FC group (n = 28, *Mean rank* = 27. 52, U = 364.50, p = 0.07, r = 0.06), we aggregated the two groups in a new "Total Clinical Sample" (TCS) for further analysis (TCS = NFC+FC, n = 56; ORS total M =17.57, SD = 8.26).

Internal consistency was calculated at intake in TCS, and at first administration (test) in NCCS. In TCS alpha value was very good (4 items solution; α = 0.83) and has not increased with items deletion. In the NCCS, the four items solution rendered the same index as in the Clinical population (TCS; α = 0.83), slightly increased with the deletion of item b to a .85. These compare to: Miller et al., (2003; .87), Janse et al., (2014; .82) and Moggia et al. (2017; 2018; .88) for the clinical population.

Construct Validity

Convergent Validity in Clinical sample (NFC)

The correlation between ORS subscales and OQ-22 subscales has been investigated where the two scales have been administered concomitantly (Table 2).

[2] http://www.statmodel2.com/discussion/messages/9/572.html?1319030818

--INSERT TABLE 2 HERE--

Spearman rho correlations between ORS Individually and OQ-22 Subjective distress subscales were significant (p <.05, moderate effect), as correlations between ORS Interpersonally and OQ-22 Interpersonal Relationship subscale (p < .05, moderate effect), and ORS Socially and OQ-22 Social Role subscale (p < .01, strong effect). ORS (T) and OQ-22 Total also correlate significantly (p < 0.05, moderate effect). Individually and Interpersonally ORS subscales correlate with each respective OQ subscales quite exclusively, while ORS Socially correlate with each OQ-subscales and ORS Total correlates with all OQ- subscales excepted IR. The strongest associations occurred between ORS Socially and OQ-22 SD (rs =-.66) and between ORS socially and OQ-22 Total (rs = -.66).

--INSERT TABLE 3 HERE--

All ORS subscales and Total scores are significantly correlated with K6 total scores (Table 3), with the lowest effect size for ORS Interpersonally (rs = -.38, moderate correlation) and the biggest effect size for ORS Socially and Overall (rs = -.65, strong correlation).

Convergent Validity in the community sample (NCCS)

--INSERT TABLE 4 HERE--

--INSERT TABLE 5 HERE--

In the Non-Clinical sample (Table 4), spearman correlations between ORS Individually and OQ-22 Subjective distress subscales were significant (p <.01, strong effect), as correlations between ORS Interpersonally and OQ-22 Interpersonal Relationship subscale (p < .01, strong effect), and ORS Socially and OQ-22 Social Role subscale (p < .01, moderate effect). ORS (T) and OQ-22 Total also correlated significantly (p < 0.01, strong effect). All ORS subscales correlate with each OQ subscale. The strongest associations occurred between ORS Individually, OQ-22 SD and OQ-22 Total (respectively r =-.63 and r = -.64) and between ORS Overall, OQ-22 SD and OQ-22 Total (r = -.68 and r = -.71) with the strongest effect size for the relation between ORS Total and OQ-22 Total (r = -71). All ORS subscales and Total scores are significantly correlated with K6 total scores (Table 5), with the lowest effect size for ORS Interpersonally (r = -.27, moderate) and the biggest effect size for ORS Overall and K6 Total (r = -.65, strong). These results compared to Miller et al. (2003, 2016), Janse et al. (2014) and Moggia et al. (2017;2018) in the magnitude of effects between ORS and other references measures and suggest that ORS, K6 and OQ-22 to a lower extent could measure related dimensions of a general distress factor.

Criterion Validity

A linear regression with BFI negative affect dimension scores (trait distress) as a predictor of ORS scores (state distress), on occasion 1, with the whole student sample (n =154), was significant (β = -.43, $R2$ =.19, p <.01). Moreover, students ORS scores at session1 in the community clinical group (CCS; *Mean Rank* = 63.60) did not significantly differ from ORS scores in the community non-clinical group (NCCS; *Mean Rank* = 80.07, U =1226.50, p = .10, r = .13). This theoretically consistent result suggests that students presenting a psychological disorder (ongoing treatment) are non-distinguishable regarding their ORS scores from non-clinical students (no treatment).

Sensitivity to Change

Ability to distinguish clinical from non-clinical population using ORS scores.

Students ORS scores at session1 in the community sample (NCCS; M = 25.65) were significantly higher than ORS scores in the total clinical group (TCS; M = 17.57, t *(97.23)* = 6.26, p <0.01, r = .42 or d= .74).

Change during treatment

Only clients in the first episode of treatment (delay between two consecutive sessions not exceeding eight weeks or 61 days), were considered. With NFC clients, last available Session ORS scores were significantly higher (M = 28.01) than client ORS scores at intake (M = 18.57, T = 123, p <0.01, r = 0.71). With FC clients, last available ORS scores were significantly higher (M = 29.15) than client ORS scores at intake (M = 17.04, T = 297, p <0.01, r = 0.86). These are arguments towards the possibility of reliably discriminate clinical, non-clinical status and progression during the treatment (change) by examining ORS scores.

Temporal Stability

Test-Retest correlation between occasion one and two in the non-clinical student sample was higher (NCCS; *mean delay* =15.92 days, *SD* = 5.31, r_s =.57, *p* <.01, 95% BCA CI [.34, .74]) than test-retest correlation between session one and two in the feedback Clinical group (FC, *mean delay* =13.08 days, *SD*=7.92, r_s =.41, p <.05, 95% BCA CI [-.05., .80]) on a comparable period of time; and higher than test-retest correlation between session one and two in the non-feedback clinical group which was nonsignificant (NFC; *mean delay* = 26.56 days, *SD* = 14.40, r_s =.31, *p* >.05, BCA CI [-.17 , .66]). As test-retest correlation coefficient is lower in clinical group compared to higher and significant in the non-clinical group, this suggests that ORS is sensitive to change and reliably assesses the change in a clinical population (FC).

Cut-off scores and RCI (Jacobson and Truax ;1991)

A cut-off score (*c)* has been computed using the Means and Standard deviations of the Non-Clinical Population (community sample) and clinical population (TCS) and was 22 (21.77). This is lower than the American cut-off (*c* = 25) and the Dutch cut-off (*c* = 24).

To give the most conservative index (Seidel et al., 2016), RCI have been computed with the community sample using the reliability of the measure (α) and the standard deviation *SD* at occasion 1 and was 9 points (8.68). It is equivalent to the Dutch RCI but higher than the American RCI (5 points). We calculated the RCI with the TCS sample and found a similar result (9.39 points).

Receiver Operating Characteristic Curve (ROC)

ROC curve has been drawn in SPSS by using the biggest sample of patients at session 1 (TCS, clinical population, *n* = 56) and the community sample (NCCS, *n* = 130). The total area

under the curve was .77, the score of 23 (22.92) rendered a fair balance between specificity and 1-sensitivity (respectively .70, .25); this is lower at one point than cut-off (*c)*, lower than the US cut-off (25; Miller et al., 2004) and lower than Dutch Cut-off (24; Janse et al., 2014) .

Session Rating Scale (SRS)

Descriptive Statistics

Table 6 presents means and standard deviations of the SRS scores at intake.

--INSERT TABLE 6 HERE--

SRS score were from client records (NFC and FC) at Session 1, with 22 scores in NFC (SRS; *M* = 34.78, *SD* = 6.52) and 28 scores in FC (SRS; *M* = 34.55, *SD* = 5.17). Client total SRS scores at session 1 in Non-Feedback Condition (NFC; *Mean Ran k*= 23.41) were not significantly different than client SRS total scores in Feedback Condition (FC; *Mean Rank* = 28.16, *U* = 249.50, *p* = 0.25, *r* = .16), which is consistent with the literature about SRS administration regarding social desirability (Reese et al., 2013). We can aggregate SRS scores at session 1 from NFC group and FC group for further analysis. SRS scores for women at session 1 in the TCS (TCS; *Mean Rank* = 27.27) did not significantly differ from men scores (*Mean Rank* = 22.06, *U* = 222, *p* =.23, *d* = 0.17).

Factorial Structure

Evidence point out SRS conveys one factor latent structure: Subscales are correlated and an EFA of the Group Session Rating scale, built on the same model of the SRS (but designated to be used in the group), rendered one factor solution (Quirk et al., 2013; Factor loading for SRS a, b, c, d subscales respectively .72, .88, .86, .90). To confirm these results, a measurement model was estimated in Mplus version 7.4 with a robust procedure (MLR; Muthén & Muthén, 2010). Fit indices indicated that a one factor solution was acceptable ($\chi^2(2$)

= 0.07, *p* =.96, *CF I*=1.00, *TLI* =1.13, *RMSEA* =0.00, 90% CI .00-.00*)*. Factor loadings for each SRS subscales were respectively (0.93, 0.87, 0.85,0.91).

Reliability

Internal consistency was calculated in TCS at session 1 (*n* = 50), and test-retest correlation between Sessions 1 and 2 (*n* =38) was computed. SRS Cronbach was excellent (α = .93) and did not increase with item deletion. Test-retest correlation between SRS scores in TCS at sessions 1 and 2 were significant (Delay in days; *M* = 17.23, *SD* = 10.95, r_s =.42, *95% BCa CI* [.28 .93] *p* <.01). This result compares with Janse et al. (2014). It is awaited that therapeutic alliance evolves over time while presenting a relative measurement invariance, so the coefficient would be moderately low.

Construct Validity

The correlations between WAI-S subscales at session 1 with the SRS subscales (NFC) reproduced in table 7, were all significant (p < .05) and ranged in effects from *rs*=.53 (moderately strong) to *rs*=.81 (clearly strong). WAI-S total scores were also significantly correlated to SRS subscales and total scores (*p* <.01, strong effect). These results suggest that SRS and WAI-S could measure related dimensions of therapeutic alliance.

---INSERT TABLE 7 HERE---

In FC group, while SRS scores were not predictive of subsequent ORS scores ($p > .05$), mean SRS scores evolved positively from session 1 to 15, excepted a drop at session 4. This suggests a trend illustrating the construction of a positive alliance with the lower mean at session 1 (34.55) and the higher at sessions 12 to 15 (mean SRS score superior to 39). In NFC group mean SRS scores evolved positively from sessions 1 to 9, excepted a drop at sessions 4 and 7, with session 1 lowest score (34.78) and session 9 highest score (40).

Criterion Validity

Due to sample size limitations, we explored the pattern of SRS scores during the five first sessions of treatment in FC group. While this trend is globally positive with a mean difference of SRS from sessions 1 to 5 of 1.71 points ($SD = 4.71$), there is also a parallel globally positive trend of scores on the ORS from sessions 1 to 6, with a mean increase of score of 11.1 points ($SD = 8.79$).

Alliance Threshold

With a total of 168 administrations in FC group, SRS mean score was at 36.91 (*SD* = 4.19). In NFC group, with 65 administrations SRS mean score was very comparable at 36.07 (*SD* = 4.61). This is an additional argument to the fact that alliance can reliably be inferred if the therapist receives the feedback or not, and an indication about the mean SRS score awaited in treatment condition. The score at the 25th percentile in NFC group for all administration was 34 (33.91) and in FC group was 36 (35.66). These go in the direction proposed by Miller & Duncan (2000, 2004) for feedback condition, with less than 24% of participants in US normative sample scoring at or below 36 (and less than 9% scoring 33 or less at every session). The score in non-feedback condition is near precedent research findings (Janse et al., 2013), and comparable to Moggia et al. (2017; *P (25*) = 35.5). The more conservative cut-off could be considered at 36, as in the original SRS validation.

Discussion

The present study is not exempt from limitations. First, the sample sizes were small (clinical population) and we used non-parametric test equivalence leading to more conservative results. Second, the CFA of the ORS yielded an improper solution (i.e., a Heywood case). As the correlation between ORS a ("individually") and ORS d ("Overall") is higher in non-feedback samples where individuals received only guidance from the scale instruction (in NFC *rs* = .90 [.73, .97], p<.01 and in NCCS *rs* = .86 [.81, 90], p<.01), and smaller in the feedback sample where the therapist helped making sense of difference (FC; *rs* = .70 [.34, .92], P<.01), this suggests participants could have interpreted more similarly "Overall" (ORS item a) and "Individual" (ORS item d) causing a multicollinearity situation. In line with Campbel et al. (2009), ORS could perhaps be used with the first 3-item total score measure; but there will be an issue of "just-identification" (i.e., CFA model).

It is important to use the preliminary cut-off and the RCI cautiously (as clinical indications in combination with the original US index with a "confidence interval frameset"). These change indices are somewhat different than original standard but are nearer to other European validation results (an awaited variation from a culture to another).

ROM are clearly promising and beginning a standard of accountable care in individual and couple therapy (Miller et al., 2006; Anker, Duncan & Sparks, 2009), the clinical use of PCOMS in French-speaking contexts could be encouraged regarding these first good psychometric results, bringing significant concrete elements of answer to the current stakes regarding access, delivering and regulation of psychological treatments in French-speaking areas.

Acknowledgements

This research did not receive funding.

The first author wants to especially thank the comity composed of clinicians from Valoris for Children and Adults of Prescott-Russell Agency in Canada (Raymond Lemay, Chantal Tassé, Madeleine Lalonde, Nicholas Cardinal), and John Deltour (Belgium) who worked to develop a unified French version of the measures.

We also want to thank Charlotte Dedenis for her English editing.

Conflict of interest statement

Dr Christophe Cazauvieilh uses PCOMS in his training activities. Dr Scott D Miller is a first author on the PCOMS scales.

REFERENCES

Anker, M. G., Duncan, B. L., & Sparks, J. A. (2009). Using Client Feedback to Improve Couple Therapy Outcomes: A Randomized Clinical Trial in a Naturalistic Setting. *Journal of Consulting and Clinical Psychology, 77*, 4, 693–704.

Arnaud, B., Malet, L., Teissedre, F., Izaute, M., Moustafa, F., Geneste, J., Schmidt, J., ... Brousse, G. (2010). Validity Study of Kessler's Psychological Distress Scales Conducted Among Patients Admitted to French Emergency Department for Alcohol Consumption-Related Disorders. *Alcoholism: Clinical and Experimental Research, 34*, 7, 1235–1245.

Bachelor, A., & Horvath, A. (1999). The therapeutic relationship. In M. A. Hubble, B. L. Duncan, & S. D. Miller (Eds.), *The heart and soul of change: What works in therapy* (pp. 133–178). Washington, DC: APA Press.

Baldwin, S. A., Wampold, B. E., & Imel, Z. E. (2007). Untangling the Alliance-Outcome Correlation: Exploring the Relative Importance of Therapist and Patient Variability in the Alliance. *Journal of Consulting and Clinical Psychology, 75*, 6, 842–852.

Bertolino, B., Bargmann, S., & Miller, S. (2012). International Center for Clinical Excellence manuals on feedback-informed treatment (FIT): Manual 1--What works in therapy: A primer. Chicago, IL: International Center for Clinical Excellence.

Bordin, E. S. (1979). The generalizability of the psychoanalytic concept of the working alliance. *Psychotherapy: Theory, Research & Practice, 16*, 3, 252–260.

Bringhurst, D. L., Watson, C. S., Miller, S. D., & Duncan, B. L. (2006). The reliability and validity of the outcome rating scale: A replication study of a brief clinical measure. *Journal of Brief Therapy, 5 (1),* 23–29.

Brown, J. Dreis, S., & Nace, D. (1999). What really makes a difference in psychotherapy outcome? Why does managed care want to know? In M. Hubble, B. Duncan, & S. Miller (Eds.), *The heart and soul of change* (pp. 389–406). Washington, DC: APA Press.

Campbell, A., & Hemsley, S. (2009). Outcome Rating Scale and Session Rating Scale in psychological practice: Clinical utility of ultra-brief measures. *Clinical Psychologist, 13,* 1, 1–9.

Cazauvieilh, C. Améliorer son efficacité clinique et dynamiser sa croissance professionnelle : des méthodes simples au service du changement. Session de poster, vendredi 12 décembre 2014, *congrès de l'Association Française de Thérapie Comportementale et Cognitive (AFTCC)*, Paris, maison de la chimie.

Cazauvieilh, C (2015). De la lutte pour ressentir à l'acceptation pour mieux agir. Une approche ACT intégrative du traitement des attaques de panique et des expériences de dépersonnalisation-déréalisation. In Seznec, J.C. (Ed.), *ACT: applications thérapeutiques*. Paris: DUNOD.

Cohen, J. (2013). *Statistical Power Analysis for the Behavioral Sciences*. Burlington: Elsevier Science.

Corbière, M., Bisson, J., Lauzon, S., & Ricard, N. (2006). Factorial validation of a French short-form of the Working Alliance Inventory. *International Journal of Methods in Psychiatric Research, 15,* 1, 36–45.

French International Test Commission (2014). International Guidelines on the Security of Tests, Examinations, and Other Assessments. [www.intestcom.org]

Flynn, R.J., Aubry, T.D., Guindon, S., Tardif, L., Viau, M. Et Gallant, A. (2003). Validation d'une version française abrégée du Outcome Questionnaire et évaluation d'un service de counselling en milieu clinique, *Canadian Journal of Program Evaluation*, 17(3), 57-74.

Hafkenscheid, Anton & Duncan, Barry & Miller, Scott. (2010). The Outcome and Session Rating Scales: A Cross-Cultural Examination of the Psychometric Properties of the Dutch Translation. *Journal of Brief Therapy*. 7 (1 & 2).

Hannan, C., Lambert, M. J., Harmon, C., Nielsen, S. L., Smart, D. W., Shimokawa, K., & Sutton, S. W. (2005). A lab test and algorithms for identifying clients at risk for treatment failure. *Journal of Clinical Psychology*, *61*, 2, 155-63.

Horvath, A. O., & Bedi, R. P. (2002). The Alliance. In J. Norcross (Ed.), *Psychotherapy Relationships That Work: Therapist Contributions and Responsiveness to Patients* (pp. 37–70). New York: Oxford University Press. Psychotherapy. Oxford, New York .

Kessler, R. G., Barker, P. R., Colpe, L. J., Epstein, J. F., Gfroerer, J. C., Hiripi, E., Howes, M. J., Zaslavsky, A. M. (2003). Screening for Serious Mental Illness in the General Population. *Archives of General Psychiatry, 60,* 2, 184.

Janse, P., Boezen-Hilberdink, L., van, D. M. K., Verbraak, M. J. P. M., & Hutschemaekers, G. J. M. (2014). Measuring Feedback From Clients: The Psychometric Properties of the Dutch Outcome Rating Scale and Session Rating Scale. *European Journal of Psychological Assessment, 30,* 2, 86-92.

John, O. P., Donahue, E. M., & Kentle, R. L. (1991). Big Five Inventory. PsycTESTS Dataset. doi:10.1037/t07550-000

Lambert, M.J., & Burlingame, G.M. (1996a). Outcome Questionnaire (OQ 45.2). Stevenson, NM: American Professional Credentialing Services.

Lambert, M.J., & Burlingame, G.M. (1996b). Outcome Questionnaire (OQ HAI-22). Stevenson, NM: American Professional Credentialing Services

Lambert, M. J. & Ogles, B. M. (1997). The effectiveness of psychotherapy supervision. In C. E. Watkins (Ed.), *Handbook of Psychotherapy Supervision* (pp. 421–446). New York: Wiley.

Lambert, M. J., & Ogles, B. M. (2004). The efficacy and effectiveness of psychotherapy. In M. J. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (5th ed., pp. 139–193). New York: Wiley.

Lambert, M. J. (2015). Progress feedback and the OQ-system: The past and the future. *Psychotherapy* (chicago, Ill.), 52, 4, 381-90.

Miller, S. D., & Duncan, B. L. (2000, 2004). The Outcome and Session Rating Scales: Administration and Scoring Manual. Chicago, IL: ISTC.

Miller, S.D., Duncan, B.L., & Johnson, L.D. (2000). The session rating scale 3.0. Chicago, IL: Authors.

Miller, S. D., Duncan, B. L., Sorrell, R., & Brown, G. S. (2005). The partners for change outcome management system. *Journal of Clinical Psychology*, 61, 2, 199.

Miller, S. D., Duncan, B. L., Brown, J., Sparks, J. A., & Claud, D. A. (2003). The outcome rating scale: A preliminary study of the reliability, validity, and feasibility of a brief visual analog measure. *Journal of Brief Therapy, 2* (2), 91–100.

Miller, S. D. and Schuckard, E (2016). Psychometrics of the ORS and SRS. Results from RCTs and Meta-analyses of Routine Outcome Monitoring & Feedback. The Available Evidence. Chicago, IL. https://www.scottdmiller.com/wp-content/uploads/2016/09/Measures-and-Feedback-2016.pdf

Nguyen, L., Blasquez, S., Bataille, B., & Chassery, C. (2012). La détresse psychologique mesurée par le score de Kessler (K6) prédit les douleurs postopératoires prolongées après chirurgie du poignet. *Canadian Journal of Anesthesia/Journal Canadien D'anesthésie*, *59*(12), 1150–1151. doi:10.1007/s12630-012-9783-8

Moggia, D., Niño-Robles, N., Miller, S.D., Feixas, G. (2017, September). Psychometric Properties of the Outcome Rating Scale (ORS) and Session Rating Scale 3.0 (SRS 3.0) in a Spanish Clinical Sample. In F. Giannone (Chair). Brief Paper Session: Quantitative & Qualitative Method. In Society for Psychotherapy Research UK & European Chapters 4th joint conference, Oxford.

Moggia, D., Niño-Robles, N., Miller, S., & Feixas, G. (2018). Psychometric Properties of the Outcome Rating Scale (ORS) in a Spanish Clinical Sample. *The Spanish Journal of Psychology*, *21*, E30. doi:10.1017/sjp.2018.32

Muthén L.K, & Muthén B.O (2010). Mplus user's guide. Los Angeles, CA: Muthén & Muthén.

Plaisant, O., Courtois, R., Réveillère, C., Mendelsohn, G. A., & John, O. P. (2010). Validation par analyse factorielle du Big Five Inventory français (BFI-Fr). Analyse convergente

avec le NEO-PI-R. *Annales Médico-Psychologiques, Revue Psychiatrique*, 168(2), 97–106. doi:10.1016/j.amp.2009.09.003

Quirk, K., Miller, S., Duncan, B., & Owen, J. (2013). Group Session Rating Scale: Preliminary psychometrics in substance abuse group interventions. *Counselling and Psychotherapy Research*, 13 (3), 194–200. doi:10.1080/14733145.2012.744425

Reese, R. J., Gillaspy, J. A., Owen, J. J., Flora, K. L., Cunningham, L. C., Archie, D., & Marsden, T. (2013). The Influence of Demand Characteristics and Social Desirability on Clients' Ratings of the Therapeutic Alliance. *Journal of Clinical Psychology*, 69 (7), 696–709. doi:10.1002/jclp.21946

Seidel, J. A., Andrews, W. P., Owen, J., Miller, S. D., & Buccino, D. L. (2017). Preliminary validation of the Rating of Outcome Scale and equivalence of ultra-brief measures of well-being. *Psychological Assessment*, 29 (1), 65–75. doi:10.1037/pas0000311

Smith, D., Crocker, L. B., Staton, C., Gillaspy, A., & Charlton, S. R. (2010, April). *Psychometric properties of the outcome rating scale in a non-clinical population*. Poster session presented at the Annual Convention of the Southwestern Psychological Association, Dallas, TX. Retrieved on the internet the19/04/2018:

Swift, J. K., & Greenberg, R. P. (2012). Premature discontinuation in adult psychotherapy: A meta-analysis. *Journal of Consulting and Clinical Psychology*, 80, 547–559.

Tracey TJ, Kokotovic AM. Factor structure of the working alliance inventory. Psych Assess 1989; 1 (3): 207–10.

Wampold, B. E. (2001). *The great psychotherapy debate: Model, methods, and findings*. Mahwah, NJ: Lawrence Erlbaum Associates.

Wampold, B. E., & Imel, Z. E. (2015). *The great psychotherapy debate: The evidence for what makes psychotherapy work.*

Wierzbicki, Michael; Pekarik, Gene. (1993). A meta-analysis of psychotherapy dropout. *Professional Psychology: Research and Practice*, Vol 24 (2), 190–195.

Table 1. Preliminary ORS normative data at session 1

	NFC		FC		NCCS	
	n =28 (21 women, 7 men)		*n* = 28 (16 women, 12 men)		*n* =130 (109 women, 21 men)	
	M	SD	M	SD	M	SD
Group	17.96	9.02	17.17	5.57	25.65	7.63
Men	21.12	10.71	19.11	6.89	22.87	9.80
Women	16.90	8.41	15.72	7.95	26.18	7.07

Note. NFC = Non-Feedback Clinical group; FC = Feedback Clinical group; NCCS = Non-Clinical Community Sample

Table 2. Correlations between the ORS subscales, OQ-22 subscales in NFC at session 1 (*n* = 28)

	OQ-22 SD	OQ-22 IR	OQ-22 SR	OQ-22 Total
ORS Individually (a)	-.42[−.69,−.002]*	ns	ns	ns
ORS Interpersonally (b)	ns	-.42 [−.71, -.03] *	ns	ns
ORS Socially (c)	-.66 [−.86, .33]**	-.39 [−.76, .08]*	-.54 [−.81,-.17]**	-.66 [−.86, -.30]**
ORS Overall (d)	-.52 [−.77, .14]**	ns	ns	-.44 [−.68, -.0.10]*
ORS Total (T)	-.48 [−.75, -.06]*	ns	-.38 [−.66, -.06]*	-.46 [−.72, -.10]*

Note. OQ-22 SD = Subjective Distress, OQ-22 IR = Interpersonal Relationship, OQ-22 SR = Social Role; Respectively reported, rs, in bracket 95% BCa CI, significance (*Correlation is significant at the .05 level (2-tailed), **Correlation is significant at the .01 level (2-tailed)), ns= Non significant)

Table 3. Correlations between the ORS subscales, K6 totals scores in NFC at session 1 (*n* = 28)

	K6 Total
ORS Individually (a)	-.55 [−.78, -.17]**
ORS Interpersonally (b)	-.38 [−.71, .03]*
ORS Socially (c)	-.65 [−.88, -.30]**
ORS Overall (d)	-.65 [−.82, -.32]**
ORS Total (T)	-.63 [−.81, -.34]**

Notes. Respectively reported, rs, in bracket 95% BCa CI, significance (*Correlation is significant at the .05 level (2-tailed), **Correlation is significant at the .01 level (2-tailed)

Table 4. Correlations between the ORS subscales, OQ-22 subscales and K6 total score in NCCS at session 1 (*n* =130)

	OQ-22 SD	OQ-22 IR	OQ-22 SR	OQ-22 Total
ORS Individually (a)	-.63 [−.74, −.50]**	-.52 [−.64, -.38]**	-.52 [−.64, -.36]**	-.65 [−.75, -.51]**
ORS interpersonally (b)	-.38 [−.53, -.20]**	-.54 [−.65, -.40] **	-.34 [−.50, -.16]**	-.46 [−.60, -.31]**
ORS Socially (c)	-.40 [−.56, -.21]**	-.51 [−.63, -.37]**	-.47 [−.62, -.31]**	-.50 [−.64, -.34]**
ORS Overall (d)	-.68 [−.78 -.55]**	-.59 [−.70, -.46] **	-.58 [−.69, -.44]**	-.71 [−.81, -.60]**
ORS Total (T)	-.64 [−.75, -.50]**	-.66 [−.76, -.55]**	-.59 [−.71, -.44]**	-.71 [−.80, -.60]**

Notes. Respectively reported, rs, in bracket 95% BCa CI, significance (*Correlation is significant at the .05 level (2-tailed), **Correlation is significant at the .01 level (2-tailed), Ns= Non significant)

Table 5. Correlations between the ORS subscales and K6 total score in NCCS at session 1 (*n* =130)

	K6 TOTAL
ORS Individually (a)	-.63 [−.73, -.50]**
ORS interpersonally (b)	-.27 [−.43, -.07]*
ORS Socially (c)	-.38 [−.55, -.19]**
ORS Overall (d)	-65 [−.75, -.50]**
ORS Total (T)	-.59 [−.70, -.44]**

Notes. Respectively reported, rs in bracket 95% BCa CI , significance (*Correlation is significant at the .05 level (2-tailed), **Correlation is significant at the .01 level (2-tailed))

Table 6. Preliminary normative data for SRS in the TCS group at session 1

	TCS *n* = 50 (17 Men, 33 Women)	
	M	*SD*
Total	34.65	5.74
Men	33.25	6.87
Women	35.37	5.03

Notes: TCS = Total Clinical Sample

Table 7. Correlations between the SRS subscales and WAI-S subscales and total score in NFC at session 1 (n = 22)

	WAI-S Bond	WAI- S Goal	WAI-S task	WAI-S total
SRS Relationship	.56 [.14, .81]**	.69 [.28 .93]**	.81 [68, .89]**	.78 [.53, .92]**
SRS Goals	.53 [.14, .80]*	.57 [.18, .87]**	.75 [.58, .85]**	.69 [.43, .85]**
SRS approach	.59 [.14, .90]**	.70 [.36, .91]**	.82 [.63 .93]**	.78 [.51, .92]**
SRS Overall	.61 [.18, .88]**	.72 [.39, .91]**	.77 [.53, .89]**	.77 [.51, .93]**
SRS Total	.62 [.20, .88]**	.70 [.40, .89]**	.80 [.63, .89]**	.78 [.51, .93]**

Notes. Respectively reported, rs, in bracket 95% BCa CI, significance (*Correlation is significant at the .05 level (2-tailed), **Correlation is significant at the .01 level (2-tailed)) .WAI-S = Working Alliance Inventory-Short form