Journal of Brief Therapy
Volume 5 • Number 1 • 2006

The Reliability and Validity of the Outcome Rating Scale: A Replication Study of a Brief Clinical Measure

Major David L. Bringhurst, MSW, LCSW, BCD
Air Force Institute of Technology/Civilian Institutions Program
University of Utah College of Social Work
Salt Lake City, Utah

Curtis W. Watson, MSW, LCSW
University of Utah College of Social Work
Salt Lake City, Utah

Scott D. Miller, Ph.D.
Barry L. Duncan, Psy.D.
Institute for the Study of Therapeutic Change
Chicago, Illinois

There is an industry-wide trend toward making outcome evaluation a routine part of therapeutic services, yet most measures are infeasible for everyday clinical use. Consequently, the Outcome Rating Scale (ORS) was developed and recently validated by its authors (Miller, Duncan, Brown, Sparks, & Claud, 2003). This article reports the findings of an independent replication study evaluating the reliability and concurrent validity of the ORS as studied in a non-clinical sample. Concurrent validity was tested by comparing the ORS with the Outcome Questionnaire 45.2 (OQ) using correlation statistics. The findings re-confirm that the ORS has high test-retest reliability, strong internal consistency, and moderate concurrent validity. Implications for clinical practice and future research are discussed.

Miller, Duncan, Brown, Sparks, and Claud (2003) point to an industry-wide trend toward making outcome evaluation a routine part of therapeutic services. They suggest that while various multi-dimensional assessments of outcome are valid and reliable, their methodological complexity, length of administration, and cost often render them infeasible for many service providers and settings. Consequently, the Outcome Rating Scale (ORS) (Miller & Duncan, 2000) was developed as an ultra-brief alternative. Miller et al. (2003) examined the instrument's psychometric properties with

both clinical and non-clinical samples, as well as the feasibility of the measure at various clinical sites. Results indicated that the ORS is a reliable and valid outcome measure that represents a balanced trade-off between the reliability and validity of longer measures, and the feasibility of this brief scale.

The present article reports the results of an independent investigation of the psychometric properties of the ORS, specifically test-retest reliability, internal consistency reliability, and concurrent validity with a non-clinical sample. The study was implemented and the data gathered and analyzed independently; to facilitate replication, the original authors were consulted about the design and then participated in the write-up and comparison of the data between the two studies. As with the original investigation, this replication study compared the ORS to the Outcome Questionnaire – 45.2 ([OQ] Lambert, Burlingame, Umphress, Hansen, Vermeersch, Clouse, & Yanchar, 1996). Results and implications for clinical practice and future research are discussed.

Methods

The Instruments: The ORS and the OQ

The ORS (Miller et al., 2003) was developed as a brief alternative to the OQ because of feasibility complaints by clinicians interfered with implementation[1] of the OQ. The ORS is a 4-item visual analogue self-report outcome measure designed for tracking client progress in every session. Each item requires the client to make a mark on a ten centimeter line where marks to left indicate more difficulties in the particular domain and marks to the right depict fewer difficulties.

Items on the ORS were tailored from three areas of client functioning assessed by the OQ; specifically, individual, relational, and social well being and functioning.

The OQ is a widely used and respected 45-item self-report scale designed for repeated measurement of client functioning through the course of therapy. The measure has high internal consistency (.93) and test-retest reliability (.84). Moderate to high validity coefficients have been reported between the scale and other well-established measures of depression, anxiety, and global adjustment. The instrument has proven particularly useful in documenting the effect of interventions due to therapy as it has been shown to be sensitive to change in a treated population while remaining stable in a non-treated population (Lambert, Burlingame, Umphress, Hansen, Vermeersch, Clouse, & Yanchar, 1996). Two studies have further documented the scale's ability to identify and improve the chances of success in cases at risk for a negative or null outcome (Lambert, Whipple, Smart, Vermeersch, Nielsen, Hawkins, 2001; Whipple, Lambert, Vermeersch, Smart, Nielsen, Hawkins, 2003).

Participants

Participants in this study were recruited from the student population at the University of Utah, College of Social Work. The non-clinical group consisted of 98 total participants made up of masters and bachelors level students. There were 67 females and 30 males (1 individual did not report their gender), ranging in age from 20 to 59. Out of 98 participants 84% (82) completed at least two administrations with 58% (57) completing

1 For a full description of the ORS, see Miller et al. (2003).

all three administrations. A further breakdown of participation rates shows that 22.4% (22) completed the 1st and 2nd administrations, 14.3% (14) participants completed the 1st administration only, and the remaining 5% (5) participants only completed the 2nd (1), 3rd (1), or the 1st and 3rd administrations (3). Attrition at the third administration was likely because this administration occurred during the week of the Thanksgiving holiday.

Procedure

Participants signed an informed consent form prior to their participation in the study. Participants received three concurrent administrations of the ORS and OQ. The sample was tested in classroom settings, with proctors administering the instruments. Retest administration used the same procedure for the 2nd and 3rd administrations over the following 1 to 3 weeks. Data were collected during the last week of October 2003 through the 3rd week of November 2003. Participant scores were excluded from overall analysis scores if they failed to complete all three administrations. A minimum of ten cases per item on the ORS (the ORS has a total of four items) was desired to ensure sample sufficiency for data testing. This minimum was met at each administration (n = 94, 79, & 60 respectively) and overall (53). The data met assumptions of normality making it suitable for parametric statistics; the Pearson product-moment correlation coefficient was used to assess concurrent validity.

Results

Normative Data

The means and standard deviations for the sample are displayed in Table 1. The mean ORS score was similar to that reported in the preliminary ORS reliability and validity study (Miller et al, 2003). Likewise the mean OQ score for this non-clinical sample was similar to the normative sample scores reported for the OQ (Lambert et al, 1996). The comparability of both the ORS and OQ mean scores provides an initial indication of confidence in the findings.

Table 1: Sample means and standard deviations for the ORS and OQ

Sample Size	Instrument	Mean	Standard Deviation
98	ORS	29.9	7.5
98	OQ	48.3	18.7

Normative data reported for the OQ (Lambert al., 1996) also suggest that individual scores do not differ due to age or gender. There were 68 females and 30 males who participated in the study (a normal ratio of females to males in a social work student population). Differences in OQ intake scores were not found between men and women ($p>.10$). Table 2 displays the means and standard deviations of the ORS scores by gender. An inspection of the table reveals a significant difference between male and female ORS intake scores ($p< .05$). This somewhat perplexing finding also occurred in the original study.

Table 2: Gender comparison of ORS means and standard deviations

	Sample Size	Mean	Standard Deviation
Males	30	27.4	9.9
Females	68	31	5.9

P<.05; two-tailed t-test comparison of ORS scores by gender

Reliability of the ORS

Internal Consistency. Internal consistency of the ORS was evaluated by using Cronbach's alpha coefficient. Cronbach's alpha was .91 for the first administration, .93 for the second, and .97 for the third. The overall alpha for all ORS administrations was .97 (n = 53; the number of participants who completed all three administrations of the ORS) and for the OQ was .98. The overall alpha for the ORS in the original study was .93.

Table 3: Cronbach's Alpha assessing internal consistency of the ORS

1st administration (n=94)	2nd administration (n=79)	3rd administration (n=60)	All administrations (n=53)
.91	.93	.97	.97

Normally an instrument with fewer than 12 items, like the ORS, would be expected to have lower internal consistency reliability than a measure with 45 items. Miller et al. (2003) explain these unusual findings: "This high degree of internal consistency reflects the fact that the four items correlate quite highly with one another, indicating that the measure can perhaps best be thought of as a global measure of distress rather than one possessing subscales for separate dimensions" (p. 95).

Test-retest Reliability. Test-retest reliability estimates were obtained through correlation testing of each administration with each following administration. These correlations statistics are found in Table 4. Surprisingly, the ORS test-retest reliability had correlations similar to those of the OQ at the same administrations. Normally the expected test-retest reliability of an ultra-brief measure would be significantly lower than that for a measure with 45 items like the OQ. When compared with Miller et al's preliminary work (2003), the ORS test-retest correlations in this sample were markedly higher, .80 compared to .66, and .81 compared to .58, when 2nd and 3rd administrations are paired from each study. Although further research is needed to explain this difference, it is likely due to the increased time between administrations in the original study.

Table 4: Test-retest reliability correlations

	2nd administration	3rd administration	Coefficient Alpha
ORS	0.80** (n=75)	0.81** (n=55)	.97
OQ	0.84** (n=78)	0.83** (n=55)	.98

(**significant at the 0.01 level, 2-tailed)

Concurrent Validity of the ORS

Concurrent validity was computed using Pearson product-moment correlations (Cohen & Cohen, 1983 not in the references) between the ORS total score and OQ total score. Table 5 displays the correlation coefficients at each administration. The first two administrations were similar while the third administration shows a higher correlation (.69). The increased correlation at the third administration may be due in part to the attrition in participation, leaving the possibility that the more consistent and reliable students remained in class (Thanksgiving week) to fill out the survey at the third administration. These correlation coefficients suggest a moderate level of concurrent validity. Miller et al (2003) showed ORS and OQ correlation coefficients of .69, .53, .54, and .56 through their four administrations respectively, suggesting a notable similarity in results. Also of note and a replication of the original study's findings in support of construct validity, the pre v. post scores of the current sample was not significant, indicating that the ORS is stable in non clinical populations.

Table 5: Pearson correlation coefficients between ORS and OQ

1st administration (n=94)	2nd administration (n=79)	3rd administration (n=60)
-0.57**	-0.56**	-0.69**

(**significant at the 0.01 level, 2-tailed)

Though modeled on the OQ, it is not reasonable to expect very high coefficients of correlation between the two measures given the shorter nature of the ORS. Nonetheless, the correlation is respectable and does provide evidence that the ORS is an ultra brief alternative for assessing global subjective distress similar to that measured by the full-scale score on the OQ.

Discussion

Outcome evaluation can be used to enlighten clinical decision-making and improve treatment effectiveness (Duncan, Miller, & Sparks, 2004; Howard, Moras, Martinovich, & Lutz, 1996). Studies of outcome feedback in psychotherapy (Lambert, Whipple, Smart,

Vermeersch, Nielsen, & Hawkins, 2001; Whipple, Lambert, Vermeersch, Smart, Nielsen, & Hawkins, 2003) have demonstrated a 65% improvement in cases most at risk for negative outcomes. Furthermore, Miller, Duncan, Brown, Sorrell, & Chalk (in press) found that the ongoing outcome feedback to clinicians doubled overall effectiveness in a sample of over 6000 clients. These dramatic results point to the importance of the development of an outcome measure that clinicians view as user-friendly as well as reliable and valid.

This article reported the results of an independent investigation of the reliability and validity of an ultra-brief outcome measure, the ORS. Although a short measure can't be expected to achieve the same specificity or breadth of information as a longer measure like the OQ, this study replicated the original validation study and found that the ORS has adequate concurrent validity, and moderate to high reliability.

It is curious that the finding that females scored significantly lower than males in the Miller et al. study was also replicated. Further research should examine this finding. Research using assorted clinical and non-clinical samples is also recommended, as well as a focus on the stability of the ORS with clinical samples prior to treatment, longitudinally, and with normal controls.

References

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlational analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum & Associates.

Duncan, B. L., Miller, S. D., Sparks, J.A. (2004). The heroic client: A revolutionary way to improve effectiveness (Revised). San Francisco: Jossey Bass.

Howard, K., Moras, K., Brill, P., Martinovich, Z., & Lutz, W. (1996). Evaluation of psychotherapy: Efficacy, effectiveness, and patient progress. *American Psychologist, 51*(10), 1059-1064.

Lambert, M.J., Burlingame, G.M., Umphress, V., Hansen, N.B., Vermeersch, D.A., Clouse, G.C., & Yanchar, S.C. (1996). The reliability and validity of the Outcome Questionnaire. *Clinical Psychology and Psychotherapy, 3*¸ 249-258

Lambert, M.J., Whipple, J., Smart, D., Vermeersch, D., Nielsen, S., & Hawkins, E. (2001). The effects of providing therapists with feedback on patient progress during psychotherapy: Are outcomes enhanced? *Psychotherapy Research, 11*(1) 49-68.

Miller, S. D., Duncan, B. L., Brown, J, Sorrell, R., & Chalk, M. (2006). Using formal client feedback to improve retention and outcome: Making ongoing real-time assessment feasible. *Journal of Brief Therapy, 5*(1).

Miller, S.D., Duncan, B.L., Brown, J., Sparks, J., & Claud, D. (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. & Duncan, B. L., (2000). *Outcome Rating Scale*. Retrieved May 29, 2003, from www.talkingcure.com; directed on its use by personal email from Scott Miller.

Whipple, J. L., Lambert, M. J. Vermeersch, D.A., Smart, D.W., Nielsen, S.L.; Hawkins, E. J. (2003). Improving the effects of psychotherapy: The use of early identification of treatment and problem-solving strategies in routine practice. *Journal of Counseling Psychology, 50*(1) 59-68.

Major David L. Bringhurst, MSW, LCSW, BCD
Air Force Institute of Technology/Civilian Institutions Program
University of Utah College of Social Work
Salt Lake City, Utah
davidandbelinda@yahoo.com

Curtis W. Watson, MSW, LCSW
University of Utah College of Social Work
Salt Lake City, Utah

Scott D. Miller, Ph.D.
Barry L. Duncan, Psy.D.
Institute for the Study of Therapeutic Change
P.O. Box 578264
Chicago, Illinois
scottdmiller@talkingcure.com

Appendix 1

Outcome Rating Scale (ORS)

Name _____Age (Yrs):____
ID# _____ Sex: M / F
Session # ____ Date: _____

Looking back over the last week (or since your last visit), including today, help us understand how you have been feeling by rating how well you have been doing in the following areas of your life, where marks to the left represent low levels and marks to the right indicate high levels.

Individually:
(Personal well-being)

I----------I

Interpersonally:
(Family, close relationships)

I----------I

Socially:
(Work, School, Friendships)

I----------I

Overall:
(General sense of well-being)

I----------I

Institute for the Study of Therapeutic Change

www.talkingcure.com

Visit www.talkingcure.com/measures.htm to download a free working version of this instrument.