Lambert, M.J., Whipple, J.L., Hawkins, E.J., Vermeersch, D.A., Nielsen, S.L. et al. (2003). Is it time for clinicians to routinely track patient outcome?: A meta-analysis. *Clinical Psychology: Science and Practice, 10*, 288–301.

Lambert, M.J., Morton, J.J., Hatfield, D., Harmon, C., Hamilton, S. et al. (2004a). *Administration and Scoring Manual for the Outcome Questionnaire-45.* SLC, UT: OQMeasures.

Lambert, M.J., Whipple, J.L., Harmon, C., Shimokawa, K., Slade, K. et al. (2004b). *Clinical Support Tools Manual.* Provo, UT: Department of Psychology, Brigham Young University.

Lueger, R.J., Lutz, W. & Howard, K.I. (2000). The predicted and observed course of psychotherapy for anxiety and mood disorders. *The Journal of Nervous & Mental Disease, 188*, 127–143.

Lueger, R.J., Howard, K. I., Martinovich Z., Lutz, W., Anderson, E.E. et al. (2001). Assessing treatment progress of individual clients using expected treatment response models. *Journal of Consulting and Clinical Psychology, 69*, 150–158.

Lutz, W., Lambert, M.J., Harmon, S.C., Tschitsaz, A., Schürch, E. et al. (2006). The probability of treatment success, failure and duration – what can be learned from empirical data to support decision making in clinical practice? *Clinical Psychology & Psychotherapy, 13*, 223–232.

Lutz, W., Leach, C., Barkham, M., Lucock, M., Stiles, W.B. et al. (2005). Predicting rate and shape of change for individual clients receiving psychological therapy: Using growth curve modeling and nearest neighbor technologies. *Journal of Consulting and Clinical Psychology, 73*, 904–913.

Spielmans, G.I., Masters, K.S. & Lambert, M.J. (2006). A comparison of rational versus empirical methods in prediction of negative psychotherapy outcome. *Clinical Psychology & Psychotherapy, 13*, 202–214.

Whipple, J.L., Lambert, M.J., Vermeersch, D.A., Smart, D.W., Nielsen, S.L. et al. (2003). Improving the effects of psychotherapy: The use of early identification of treatment failure and problem solving strategies in routine practice. *Journal of Counseling Psychology, 58*, 59–68.

Yalom, I.D. & Lieberman, M.A. (1971). A study of encounter group casualties. *Archives of General Psychiatry, 25*, 16–30.

7

Treatment Outcome Package (TOP) – Development and use in Naturalistic Settings

David Kraus[1] and Louis G. Castonguay[2]

[1]Behavioral Health Laboratories, Marlborough, MA, USA, [2]Pennsylvania State University, USA

Introduction

This chapter describes the development of a widely used outcome tool designed specifically for naturalistic, real-world treatment settings. From large networks like Blue Cross and Blue Shield of Massachusetts, to solo-practice clinicians in Canada and other countries, TOP (the Treatment Outcome Package) has emerged as a popular outcome tool for a wide variety of provider groups. By the beginning of 2007, TOP, created and processed by Behavioral Health Labs (BHL), had been used by more than 30,000 clinicians to assess more than 600,000 clients, their treatment needs and their progress towards defined treatment plan goals.

The TOP features a number of characteristics that makes it suitable to both evidence-based practice and clinically relevant research. For example, based on Lambert-style alert methodology (see Chapter 6), it provides individualized client feedback to warn clinicians of potentially poor outcomes. Aggregate report feedback is risk adjusted and benchmarked with ties to empirically based quality improvement suggestions, helping providers to maximize the type of feedback that can accelerate treatment effectiveness. With an anonymous (patient de-identified) benchmarking database exceeding a million clients, the opportunities for research and fine-tuned benchmarking are remarkable. With such a massive and centralized database, BHL is able to provide detailed,

Developing and Delivering Practice-Based Evidence By Michael Barkham, Gillian E Hardy, and John Mellor-Clark

risk-adjusted benchmarking that allows clinicians to discover their strengths and weaknesses compared to their peers who treat similar populations.

To us, this is the ultimate example of evidence-based practice – if your benchmarked results are excellent, you have well-measured evidence assuring you to keep doing what you have been doing – good work. If your results in a particular area are subpar, then you should consider incorporating practice-based changes like empirically validated, evidence-based treatments.

Using quartile analyses, the TOP system will compare your best-treated and worst-treated populations, painting a clear picture of the demographic and clinical characteristics of the clients who could benefit from your practice improvement (e.g. consultation, focused continuing education and reading professional publications). Then, with a library of catalogued evidence-based principles and therapies tied to each TOP outcome domain, and within several minutes of clicking a button to access the benchmarking and quartile analyses, the treatment provider can know what concrete actions to consider taking. Below, we detail the vision, development and application of this system.

Vision and philosophy

In the early 1990s, as the first author was planning to leave his post as a team leader of an inpatient adolescent program to join a group practice outside Boston, Hillary Clinton was talking about national healthcare reform. Although her efforts eventually failed, she gave extraordinary momentum to the concept of accountability in healthcare. The concept was bold yet simple – all healthcare providers should be able to document that, on average, their patients were benefiting from the services provided. No longer would we be paid for delivering services on blind faith.

In an ideal world, the concept of measuring outcomes is hard to argue with. As providers, we probably should not be arguing against the measurement of quality, even if it is difficult and it may never be error free. Resisting measurement looks like we have something to hide. Nevertheless, the concept, while supported whole-heartedly by my brain, sent shivers up my spine. How would this newly minted, solo-practice provider compare to all the experienced therapists on managed care panels? Would I be first on the chopping block if the industry moved forward with its first, half-baked plans to throw a third of us into the ocean if our outcomes were subpar (Geigle & Jones, 1990)?

It was this anxiety that started the snowball rolling, leading to the development of the TOP. The first author's initial, counter-phobic response was to research the field and look for the best outcome system for his group practice to adopt. We were not willing to wait for Hillary or local managed care companies to decide how to measure our quality behind our backs while we sat back idly

waiting for their verdicts. We wanted to know where we stood before they did, and afford ourselves the chance of taking action if it looked like action was necessary.

The snowball grew as these trips to the library failed to find a world-class outcome system for real-world providers. From practical issues of cost, to the scientific problems of risk adjustment, basic issues could not be answered to our satisfaction. For example, if insurance companies were really planning to compare our results to others, they should not do this based on raw outcome data. This was particularly obvious for the first author who, at the time, specialized in treating adolescents and young adults with multigenerational issues of abuse and trauma and whose outcomes were more likely to differ from those specializing in treating younger children, or the worried well. However, we saw no system or methodology that took these case-mix variables seriously, and certainly no outcome tool that did either.

Frankly, there were too many concerns to list here. Rather than crying foul and waging war against managed care, we decided to take action and begin the development of an outcome system built by, and for, providers. The five guiding principles that launched the development of TOP still guide our vision today: utility, cost, timing, risk adjustment and benchmarking.

Utility

We believe that the entire philosophy and approach to most outcome movements have been off-target. We certainly believe that the pressure for accountability is here to stay; however, it should neither be the single, nor the most important use of outcome data. As hinted to above, the entire process got off on the wrong foot when the major healthcare players gathered in the late 1980s to discuss the use of health outcomes (Geigle & Jones, 1990). Their meeting had overwhelmingly punitive tones. For example, the consensus, number-one use of outcome data was to profile clinicians on outcomes and eliminate those with 'documented poor quality'. With such approaches, there is little reason to expect clinician buy-in, or evidence-based use of the data; clinicians would be running scared.

We believe the principal focus of outcomes should be to guide and assist the psychotherapist in planning the treatment process. Such a tool should never prescribe a certain intervention, but provide the clinician with information, tailored to the patient's assessment and condition, about the relative success of various treatment options, and outline current advances in standard care by pointing to evidence-based interventions. By properly guiding clinicians, a system of outcomes management can facilitate communication between the patient and clinician while helping to identify budding problems before they become serious. Such a system is much more likely to be embraced by clinicians

because it can inform and potentially improve the therapeutic process, rather than just evaluating and judging it.

Cost

To be used in clinical practice, we believe that outcomes management tools need to be free of royalty charges. Data processing services should be offered at cost to encourage use and capitalize on economies of scale. In line with this philosophy, BHL does not charge any royalty fees for the use of TOP. Unless you make other arrangements, however, you do need to use their service bureau to process the data. The advantages of this requirement are highlighted below (see section below on Benchmarking).

Timing

The major reason previous generation outcome projects failed is because of data processing. From Georgia to Washington State there are countless examples of massive amounts of data being dumped into black holes. Needless to say, it is impossible to sustain a project that cannot deliver useful results to its key participants – the patient and the psychotherapist. Whether the data is processed electronically or on paper, the BHL TOP system is designed to return useful results with this urgency in mind. Paper processing is obviously the most challenging:

- After the patient completes a TOP, the form is faxed to BHL's central computer system. There, it never touches paper again. A TIFF file image (the computer graphic file generated by your fax machine) is transferred to three data processing engines that translate the images into data.
- A human verifier looks over every form and makes sure the computers have correctly processed the information. The data is then transferred to the data warehouse where it is scored, compared to general population norms and any historical patient records, and a report is generated.
- These reports are returned via fax or e-mail to the clinician with an average return time (from hitting send on your fax machine) of 14 minutes.
- As an alternative to a fax-based system, BHL also has an electronic/web system where the results are returned within three seconds.

BHL also provides toll-free customer service, a training video and extensive documentation, making startup simple. By offloading the time-consuming process of warehousing and scoring reports, clinicians can stay focused on what they do best – treatment.

Similar to psychological testing, outcome assessment data and their reports need to be fed back to clinicians in a timely manner so that the results can

be integrated into treatment planning, evaluation and diagnostic formulations. Only by delivering reports that facilitate the treatment process can a system meet clinicians' needs and win their buy-in.

Risk adjustment

Variables that are beyond the control of the therapeutic process, but nonetheless influence the outcome, are defined as case-mix variables (Goldfield, 1999). Naturalistic research typically lacks the experimental controls used in efficacy research to militate against the need to statistically control for (or measure) these case-mix variables. Without measuring and controlling for such variables, comparing or benchmarking naturalistic datasets can be quite misleading. Hsu (1989) has shown that even with randomization, when the samples are small, the chances of a 'nuisance' case-mix (e.g. AIDS) variable being disproportionately distributed across groups, is not only common, but very likely (in some cases exceeding 90%). Therefore, without extensive case-mix data, results have limited administrative value in real-world settings. These data need to be used to disaggregate and/or statistically adjust outcome data to produce fair and accurate benchmarking. Unless your clients' levels of life stress, or co-morbid medical conditions were measured and adjusted for, we believe your benchmarked outcomes would likely lead to misleading results.

Benchmarking

Outcome data is of little use if we do not have a reference group for comparison. Furthermore, since most payers probably will not blindly trust outcome data presented to them by providers (see Bilbrey & Bilbrey, 1995), we made a strategic decision to take on the difficult work of becoming a neutral, third-party data-processing organization. Like the consumer reports of the behavioural health field, BHL can certify outcome results as free from bias and fraud while offering the largest possible benchmarking reference database.

Whether it is looking for a sample of eating-disorder clients treated in residential treatment settings, or a sample of solo-practice providers specializing in treating sexual dysfunction, BHL is able to offer this level of specificity when creating a reference sample for your results. From here, with sophisticated risk adjustment, it is much easier to know where our relative strengths and weaknesses really lie.

Development

The TOP and its supporting infrastructure are designed to bring to the forefront the positive and beneficial aspects of outcomes management. For a kinder or

Table 7.1 Core Battery Conference criteria for a universal core battery

Not bound to specific theories
Appropriate across all diagnostic groups
Must measure subjective distress
Must measure symptomatic states
Must measure social and interpersonal functioning
Must have clear and standardized administration and scoring
Norms to help discriminate between patients and non-patients
Ability to distinguish clients from general population
Internal consistency and test–retest reliability
Construct and external validity
Sensitive to change
Easy to use
Efficiency and feasibility in clinical settings
Ease of use by clinicians and relevance to clinical needs
Ability to track multiple administrations
Reflect categorical and dimensional data
Ability to gather data from multiple sources

Source: Horowitz, L.M. et al. (eds.), *Measuring patient change in mood, anxiety, and personality disorders: Toward a core battery.* Washington, DC: American Psychological Association Press, 1997

friendlier outcomes management system (Kraus, Castonguay & Wolfe, 2006) to be clinically helpful, however, it needs to rest on solid psychometric properties. In this next section we will discuss the empirical basis of the TOP's development, which parallels the recommendations of the 1994 Core Battery Conference that was organized by the Society for Psychotherapy Research and the American Psychological Association (Horowitz, Lambert & Strupp, 1997). These recommendations are listed in Table 7.1.

As a universal core battery, the TOP is not tied to any specific theoretical orientation and measures many categories within symptom, functional and quality-of-life domains. The current version of the TOP is in its fourth incarnation with 48–58 questions, depending upon the age version (child, adolescent and adult).

Construction of the TOP

Initial development of the first version of TOP consisted of the first author generating more than 250 atheoretical items that spanned diagnostic symptoms and functional areas identified in the Diagnostic and Statistical Manual of Mental Disorders-IV (DSM-IV; American Psychiatric Association, 1994). All DSM-IV Axis I diagnostic symptoms were reviewed and those symptoms that the first author thought clients could reliably rate on a self-report measure

were formulated into questions. Many assessment tools were also reviewed for item inclusion, but most were based on theoretical constructs inconsistent with DSM-IV symptomatology.

These questions were then presented to other clinicians for edit and review. They made suggestions for modifications and deletions, based on relative importance and clarity of items. Clients were administered initial versions of the questionnaires and asked for feedback as well. Questions were reworded based on feedback, and items that were less important or appeared to measure a similar symptom were eliminated. The tool was then revised and re-introduced for feedback. Such expert clinical and client review in the development process ensured adequate face validity. Subsequent versions of TOP were also used in clinical practice and the data factor analysed with increasingly robust sample sizes. The instrument presented here is the result of four iterations of this process.

The current version of the TOP is a battery of distinct modules that can be administered all together or in combinations as needed. The various modules of the TOP include:

- Chief complaints
- Demographics
- Treatment utilization and provider characteristics
- Co-morbid medical conditions and medical utilization
- Assessment of life stress
- Substance abuse
- Treatment satisfaction
- Functioning
- Quality of life/Subjective distress
- Mental health symptoms.

Psychometric overview

Factor structure

The development of stable and clinically useful subscale structures is a long-term process that is often short-changed in tool development. Nevertheless, it is a critically important step in creating a tool that is rich in reliable, valid and clinically useful information. A 13-year path aimed at creating a robust and parsimonious questionnaire with subscales that meet the criteria of the Core Battery Conference led to the current version of the TOP.

In the latest iteration, the 93 mental health symptom, functional and quality-of-life items from the third TOP version were administered to a large sample

of newly admitted psychiatric clients. Participants were instructed to rate each question in relation to 'How much of the time during the last month you have . . .'. All questions were answered on a 6-point Likert frequency scale: 1 (*All*), 2 (*Most*), 3 (*A lot*), 4 (*Some*), 5 (*A little*), 6 (*None*).

The sample consisted of 19,801 adult patients treated in 383 different behavioural health services across the United States who completed all questions of the TOP at intake, as part of standard treatment. The sample was split into five random subsamples as a cross-validation strategy.

Sample 1 was used to develop a baseline factor model. Responses to the 93 items were correlated, and the resulting matrix was submitted to principal-components analysis (PCA) followed by correlated (Direct Oblimin) rotations. The optimal number of factors to be retained was determined by the criterion of eigenvalue greater than one supplemented by the scree test and the criterion of interpretability (Cattell, 1966; Tabachnick & Fidell, 1996). Items that did not load greater than 0.45 on at least one factor, and factors with fewer than three items were trimmed from the model.

Sample 2 was then used to develop a baseline measure of acceptability in a Confirmatory Factor Analysis (CFA) and revise the model using fit diagnostics in AMOS 4.0 (Arbuckle & Wothke, 1999). Goodness of fit was evaluated using the root mean square error of approximation (RMSEA) and its 90% confidence interval (90% CI; cf. MacCallum, Browne & Sugawara, 1996), comparative fit index (CFI), and the Tucker-Lewis index (TLI). Acceptable model fit was defined by the following criteria: RMSEA (<0.08, 90% CI<0.08), CFI (>0.90), and TLI (>0.90). Multiple indices were used because they provide different information about model fit (i.e. absolute fit, fit adjusting for model parsimony, fit relative to a null model); used together these indices provide a more conservative and reliable test of the solution (Jaccard & Wan, 1996). Few outcome tools' substructures have passed this state-of-the-art method of confirming a tools' factor structure.

For the TOP, most of the revised models were nested; in these situations, comparative fit was evaluated by χ^2 differences tests (χ^2_{diff}) and the interpretability of the solution. The final model that resulted from Sample 2 exploratory procedures was then comparatively evaluated in three independent CFAs (Samples 3–5) using the criteria above with excellent results (exceeding all pre-defined criteria as demonstrated in Table 7.2). Taken together, these analyses provide strong support for the stability and strength of the TOP factors.

Test–retest reliability

In order to assess the test–retest reliability, 53 behavioural health clients recruited by four community mental health centres completed the Treatment Outcome Package one week apart while they were waiting for outpatient

Table 7.2 Confirmatory factor analysis fit statistics

Confirmatory Factor Analysis	*Description*	*N*	*DF*	*Tucker-Lewis Index*	*Comparative Fit Index*	*Root Mean Square Error of Approximation Indices*	
						RMSEA	*RMSEA Upper*
Sample 2 initial	Derived from EFA model	3960	1218	0.898	0.906	0.045	0.046
Sample 2 final	Modified model	3960	1007	0.945	0.951	0.033	0.034
Sample 3	Confirmatory Analysis 1	3960	1007	0.940	0.946	0.035	0.036
Sample 4	Confirmatory Analysis 2	3960	1007	0.942	0.948	0.034	0.035
Sample 5	Confirmatory Analysis 3	3961	1007	0.940	0.947	0.035	0.036

treatment to begin. The stability of the TOP over time was assessed by computing intraclass correlation coefficients using a one-way random model. Except for the subscale Mania, all reliabilities for subscales (factors presented in Study 1) were excellent (see Table 7.3), ranging from 0.87 to 0.94. The subscale Mania's reliability was acceptable, but considerably lower at 0.76. This is due to the bi-modal distribution of mania items like: 'Feeling on top of the world'. These items do not have a linear relationship to health as feeling on top of the world all of the time might be an indication of mania, while never feeling on top of the world might be a sign of depression.

Discriminant and convergent validity

An important step in the establishment of the validity of a measure is the testing of whether it correlates highly with other variables with which it should theoretically correlate (convergent validity), and whether it does not correlate significantly with variables from which it should differ (discriminant validity). For the purpose of examining convergent and divergent validity, 312 participants completed the TOP and one or more validity questionnaires, outlined as follows: 110 completed the BASIS 32 (51 general population, 23 outpatient and 36 inpatient), 80 completed the SF-36 (43 general population, 3 outpatient and 34 inpatient), and 69 completed the BSI, BDI and MMPI-2 (69 outpatient).

Results provided evidence to the effect that TOP factors are measuring the constructs they were intended to measure. The TOP depression scale, for instance, correlated 0.92 with the Beck Depression Inventory (for more details, see Kraus, Seligman & Jordan, 2005).

Floor and ceiling effects

For an outcome tool to be widely applicable (especially for populations like the seriously and persistently mentally ill) it must accurately measure the full spectrum of the construct, including its extremes. Using a measure that fails to capture extreme level of psychopathology would be comparable to the use of a basal body thermometer (with a built-in ceiling of only 102 degrees) to study air temperature in the desert. On a string of hot summer days, one might conclude that the temperature never changes and stays at 102 degrees.

For a psychiatric patient who scores at the ceiling of the tool but actually has much more severe symptomatology, the patient could make considerable progress in treatment, but still be measured at the ceiling on follow-up. Incorrectly concluding that a client is not making clinically meaningful changes can lead to poor administrative and clinical decisions. The SF-36, for example,

Table 7.3 Subscale intercorrelations

FACTOR	*DESCRIPTION*	*DEPRS*	*VIOLN*	*SCONF*	*LIFEQ*	*SLEEP*	*SEXFN*	*WORKF*	*PSYCS*	*PANIC*	*MANIC*	*Alpha*	*Intraclass test–retest*
DEPRS	Depression											0.93	0.93
VIOLN	Violence	0.33										0.81	0.88
SCONF	Social Conflict	0.55	0.33									0.72	0.93
LIFEQ	Quality of Life	−0.78	−0.24	−0.45								0.85	0.93
SLEEP	Sleep Functioning	0.64	0.26	0.41	−0.50							0.86	0.94
SEXFN	Sexual Functioning	0.51	0.21	0.38	−0.41	0.36						0.69	0.92
WORKF	Work Functioning	0.55	0.43	0.53	−0.41	0.37	0.34					0.72	0.90
PSYCS	Psychosis	0.66	0.55	0.42	−0.46	0.51	0.42	0.50				0.69	0.87
PANIC	Panic	0.73	0.33	0.43	−0.52	0.59	0.43	0.46	0.67			0.83	0.88
MANIC	Mania	−0.26	0.11	−0.09	0.37	−0.12	−0.09	0.01	0.05	0.04		0.53	0.76
SUICD	Suicidality	0.44	0.44	0.26	−0.33	0.27	0.23	0.36	0.61	0.38	−0.02	0.78	0.90

has been shown to have significant ceiling effects in clinical samples (Nelson et al., 1995), suggesting that the tool has limited applicability to the Medicaid population for which it was being tested. For the TOP to be reliable and valid, it must demonstrate that it can measure the full range of pathology.

This important issue was examined using the TOP administrations for all adult clients from a diverse array of service settings that contracted with Behavioral Health Laboratories between the years of 1996 and 2003 to process and analyse their clinical outcome data (N = 216, 642). The dataset was analysed for frequency counts of clients who scored at either the theoretical maximum or minimum score of each TOP scale (Table 7.4). TOP scores are presented in Z-scores, standardized by using general population means and standard deviations. All scales are oriented so that higher scores indicate more symptoms or poorer functioning.

Analysis of the TOP revealed no substantial ceiling effects on any TOP scales, suggesting that the TOP sufficiently measures into the clinically severe extremes of these constructs. Furthermore, each TOP subscale measures at least a half to more than two standard deviations into the 'healthy' tails of its construct. Therefore, from this very large clinical sample it is reasonable to conclude that each TOP scale measures the full range of clinical severity and, as such, represents a substantial improvement over the widely used naturalistic outcome tools reported previously.

Sensitivity to change

The more accurately an outcome measure is able to measure important (even subtle) changes in clinical status, the more useful it is as an outcome tool. Unfortunately, many state governments and private payers have mandated the use of outcome tools that have inadequate sensitivity to change, costing all involved extensive time and wasted resources, only to have the project abandoned after the data are unable to demonstrate differences in provider outcomes. For example, the functional scales of the Ohio Youth Scales are not showing change in functional status in treatment (Ogles et al., 2000).

To examine its sensitivity to change, 20,098 adult behavioural health clients were administered the TOP at the start of treatment and later after several therapy sessions. For each TOP subscale, within group Cohen's d effect sizes were calculated comparing subscale scores at first TOP administration to subscale scores at the second TOP administration. In addition, a reliable change index was calculated for each TOP factor using procedures outlined in Jacobson, Roberts, Berns and McGlinchey (1999). The reliable change index can be used to determine if the change an individual client makes is beyond the measurement error of the instrument. We used the indices to classify each client as having made reliable improvement (or reliable worsening), or not, on each TOP subscale.

Table 7.4 Floor and ceiling effects

Factor	*Theoretical Minimum*	*Theoretical Maximum*	*Number of clients at Minimum*	*Number of clients at Maximum*	*Total Sample Size (N)*	*Percentage of clients at Minimum*	*Percentage of clients at Maximum*
DEPRS	−1.67	4.63	7,519	2,406	212,589	3.5	1.1
VIOLN	−0.44	15.44	121,625	978	205,932	59.1	0.5
SCONF	−1.44	2.87	11,606	726	145,695	8.0	0.5
LIFEQ	−2.34	5.05	4,430	6,210	156,738	2.8	4.0
SLEEP	−1.43	3.73	23,106	5,907	206,677	11.2	2.9
SEXFN	−1.15	3.79	48,905	1,264	150,576	32.5	0.8
WORKF	−1.54	5.95	22,081	163	152,511	14.5	0.1
PSYCS	−0.93	13.23	33,900	339	202,306	16.8	0.2
PANIC	−1.13	7.59	30,444	1,153	212,474	14.3	0.5
MANIC	−1.57	4.75	16,779	474	211,802	7.9	0.2
SUICD	−0.51	15.57	58,388	702	211,836	27.6	0.3

In addition, the same indices were used to calculate the number of clients who showed reliable improvement (or reliable worsening) on at least one TOP subscale.

For each TOP subscale, Table 7.5 presents sample size, mean and standard deviation of first and second TOP administrations, within-group Cohen's d effect size, and the percentage of clients who showed reliable improvement or worsening. With an average of only seven treatment sessions, Cohen's d effect sizes ranged from 0.16 (Mania) to 0.53 (Depression). Most TOP measures showed reliable improvement for at least a quarter of participants, and 91% of clients showed reliable improvement on at least one TOP subscale. As one might expect, the functional domains (Social Conflict, Work and Sex) tended to show less change than the symptom domains.

Criterion validity

The criterion validity of the TOP was examined in a study involving 94 members of the general population. Binary logistic regression was applied to each set of the 94 general population participants and a matched sample from the clinical population. These analyses combined all of the TOP measures into a binary stepwise logistic regression to determine the most parsimonious collection of subscales accounting for independent prediction of client vs. general population status. In this type of analysis, independent variables are entered into the equation one at a time based on which variable will add the most to the regression equation. The 10 available TOP scales (Depression, Violence, Quality of Life, Sleep, Sexual Functioning, Work Functioning, Psychosis, Mania, Panic and Suicide) served as the independent variables and client/general population status served as the dependent variable.

To explore the amount of variance accounted for in client/general population status by the six significant predictors in Analysis 1, we employed the Nagelkerke R^2 test (Nagelkerke, 1991). Quality of Life accounted for 28% of the variance in client/general population status, Psychosis accounted for another 6%, Mania accounted for another 8%, Suicidality accounted for another 5%, Work Functioning accounted for another 3%, and Sexual Functioning accounted for another 4%. Thus, together these six variables accounted for 54% of the variance in predicting client/general population status.

Ten separate samples and analyses were performed. The percentage of participants correctly classified as being from a client or general population sample ranged from 80% to 89%, with an average of 84%. Nagelkerke R^2 for the complete models ranged from 0.54 to 0.77 with a mean of 0.65. In addition, the variables that were significant predictors of client/general population status were fairly consistent across the ten analyses. In ten of the analyses, Quality of Life and Mania were significant predictors; in nine of the analyses Sexual

Table 7.5 Reliable improvement and worsening

Variable	*N*	*Initial Mean*	*Follow-up Mean*	*Initial SD*	*Follow-up SD*	*Cohen's d*	*Percentage of clients showing reliable improvement*	*Percentage of clients showing reliable worsening*
DEPRS	19,660	1.34	0.48	1.68	1.55	0.53	54	14
VIOLN	18,765	1.25	0.68	2.97	2.37	0.21	31	17
SCONF	8,047	0.28	−0.04	1.08	1.01	0.31	38	18
LIFEQ	10,039	2.19	1.44	1.83	1.81	0.41	52	21
SLEEP	18,869	0.68	0.16	1.46	1.32	0.37	47	20
SEXFN	9,407	−0.12	−0.31	1.12	1.04	0.18	25	15
WORKF	9,600	0.30	−0.10	1.44	1.29	0.29	39	20
PSYCS	18,320	2.02	1.14	2.85	2.42	0.33	44	18
PANIC	19,701	1.36	0.75	1.93	1.73	0.33	41	17
MANIC	19,561	−0.31	−0.47	1.00	0.96	0.16	10	6
SUICD	19,562	2.38	1.14	3.69	2.80	0.38	42	14

Functioning was a significant predictor; in eight of the analyses Psychosis was a significant predictor, and in six of the analyses Work Functioning and Panic were significant predictors. Other significant predictors included Suicidality (three analyses), Violence (three analyses), Depression (two analyses), and Sleep (one analysis). The most important predictor of client/general population status for each of the ten analyses was Quality of Life. This result, in and of itself, is important, as most of the frequently used outcome measures do not assess for the quality of life. By only focusing on level of distress and impairment, these instruments may fail to capture issues of meaning, purpose and/or satisfaction about oneself and his/her life. Such issues of human existence may well be a determinant in leading some people to decide to go into therapy. They are also likely to reflect some of the benefits that both client and therapist expect (implicitly or explicitly) from therapy. Similarly, this finding is consistent with Frank's (1976) demoralization hypothesis, which states that most clients do not enter therapy solely because of psychiatric symptoms. In addition to such psychological problems, clients come to therapy in a state of mind that is characterized by feelings of alienation, isolation, hopelessness, helplessness, impotence and/or a sense of meaninglessness. Such experiences, needless to say, are likely to impact on, and/or reflect, one's view of the quality of his/her life.

The results demonstrate that the TOP has some ability to discriminate between clients and members of the general population with an average correct classification rate of 84%. The consistency across the 10 separate analyses lends credence to these results. It is possible that the analyses could be further improved by adding several other scales to the analysis. The Social Conflict and Substance Abuse subscales of the TOP were not available for this analysis because these scales have been revised since the general population sample was collected.

Current applications

With more than 13 years of experience and a database that is doubling in size every few years, BHL and the TOP have a wealth of developed applications. In this section, we will highlight several of them.

Patient reports that inform

TOP questions have high-face validity to patients and psychotherapists alike. Questions are easy to read (5th grade level) and are related to DSM symptoms that are key to an initial interview (e.g. 'felt little or no interest in most things'). Years of exploratory and confirmatory factor analytic work on the TOP items

reduced the number of questions to the three-to-five most powerful questions in a broad array of clinically useful domains. For the adult version, TOP domains include: Depression, Panic, Mania, Psychosis, Sleep, Sex, Work, Quality of Life, Substance Abuse, Suicide and Violence. In contrast with outcome tools that address only one or a few dimensions of functioning, the TOP patient reports provide a wealth of clinically useful assessment data that can be easily integrated into treatment planning. Results are reported as normalized Z-scores that represent their deviation from population norms. In addition to clinical domains (along Axis I) mentioned above, diagnostic considerations are reported for Axis III (medical considerations) and IV (life stress).

BHL is also finalizing a pre-filled, yet modifiable treatment plan (based on TOP responses) that is returned along with the standard TOP report, helping the therapist save time in developing an individualized course of treatment.

With assessment of dimensions like medical utilization, prior treatments, life stress and co-morbid medical conditions, the TOP also helps paint a full picture of the patient. Clinicians can give a new patient an access code to go online and complete the TOP before the appointment. This allows the clinician to get an excellent picture of the patient's perspective of his/her difficulties before actually conducting the initial interview.

Links to the research

As an empirically anchored outcomes management system, BHL has incorporated many features of the current evidence-based movement within the development of the TOP. For example, each of the TOP domains has been linked to a library of evidence-based practices, guidelines and research findings that should help clinicians find the most effective treatments for patients with different TOP profiles. For example, if a patient scores very high on the Depression Scale, this TOP library integrates findings compiled by Castonguay and Beutler (2005) and other sources into an easy-to-read summary of state-of-the-art treatments.

Moreover, building on the seminal research of Michael Lambert – who has single-handedly demonstrated that outcomes management makes us all more effective clinicians – the TOP provides early warnings if treatment appears to be heading in an unhelpful direction. Whether it might be by suggesting strategies to repair alliance ruptures, or the need to incorporate adjunctive interventions to increase client's social support, the evidence-based interventions that BHL will soon be able to suggest to clinicians is likely to help them reduce the number of patients categorized as 'negative responders'.

The rich database accumulated by BHL is also providing opportunities to study new ways of administering items to patients that could lead to further development of the TOP. Recent developments in item response theory and computerized adaptive testing indicate that clinically reliable and meaningful

results can be obtained from responses to only a few items. The BHL database of TOP results is being analysed to identify those sets of items that have the optimal specificity and clinical 'bandwidth' to evaluate symptoms and change. The results of analyses is likely to lead to a shorter version of the TOP, which could potentially increase its clinical usefulness.

It should also be mentioned that some of recommendations for the use of the TOP in routine clinical practice have also been influenced by empirical research. Among these recommendations is the review of initial reports with patients in order to facilitate an informed discussion of the priorities and challenges of their treatment. This practice guideline is based not only on clinical experience, but also on the findings of six controlled studies showing that patients are more honest about shame-based issues on questionnaires than they are in face-to-face initial evaluations (Carr & Ghosh, 1983; Erdman, Klein & Greist, 1985; Hile & Adkins, 1997; Lucas, 1977; Searles et al., 1995; Turner et al., 1998). As such, integrating an outcome questionnaire is likely to open beneficial channels of communication between clients and therapists.

Enlightening aggregate data

Every month, BHL sends an aggregate report that summarizes the changes of a psychotherapist's average patient from intake and plots the changes their patients report over the course of treatment. Since more than 91% of patients report clinically and statistically significant change in at least one dimension of functioning, the TOP can provide very rewarding statistics to help psychotherapists guide their work.

In addition, BHL provides psychotherapists with unlimited access to its enormous benchmarking database. Psychotherapists can profile the types of patients with whom they work best and those patients with whom they need to improve their clinical skills. We have tried to use this database to identify the proverbial 'supershrink', the ideal psychotherapist who is well above average on everything. However, the data suggests that there is no such psychotherapist – we all have our strengths and weaknesses. A more realistic goal is for all clinicians to monitor their personal strengths and weaknesses by comparing their clinical outcomes with other professionals using a standardized instrument. Designed by, and for, clinicians TOP and its extensive benchmarking database is designed to facilitate this quest for learning and professional development.

References

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

Arbuckle, J. & Wothke, W. (1999). *AMOS 4.0 User's Guide.* Chicago: Smallwaters Corporation, Inc.

Bilbrey, J. & Bilbrey, P. (1995). Judging, trusting, and utilizing outcomes data: A survey of behavioral healthcare payors. *Behavioral Healthcare Tomorrow, 4,* 62–65.

Carr, A.C. & Ghosh, A. (1983). Response of phobic patients to direct computer assessment. *British Journal of Psychiatry, 142,* 60–65.

Castonguay, L.G. & Beutler, L.E. (eds.) (2005). *Principles of therapeutic change that work.* New York, NY: Oxford University Press.

Cattell, R.B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1,* 245–276.

Erdman, H.P., Klein, M. & Greist, J.H. (1985). Direct patient computer interviewing. *Journal of Consulting and Clinical Psychology, 53,* 760–773.

Frank, J.D. (1976). Restoration of morale and behavior change. In A. Burton (ed.), *What makes behavior change possible?* New York: Brunner/Mazel.

Geigle, R. & Jones, S.B. (1990). Outcomes measurement: A report from the front. *Inquiry, 27,* 7–13.

Goldfield, N. (ed.) (1999). *Physician profiling and risk adjustment.* Frederick, MD: Aspen Publishers.

Hile, M.G. & Adkins, R.E. (1997). Do substance abuse and mental health clients prefer automated assessments? *Behavior Research Methods, Instruments & Computers, 29,* 146–150.

Horowitz, L.M., Lambert, M.J. & Strupp, H.H. (eds.) (1997). *Measuring patient change in mood, anxiety, and personality disorders: Toward a core battery.* Washington, DC: American Psychological Association Press.

Hsu, L.M. (1989). Random sampling, randomization, and equivalence of contrasted groups in psychotherapy outcome research. *Journal of Consulting and Clinical Psychology, 57,* 131–137.

Jaccard, J. & Wan, C.K. (1996). *LISREL approaches to interaction effects in multiple regression.* Thousand Oaks, CA: Sage Publications.

Jacobson, N.S., Roberts, L.J., Berns, S.B. & McGlinchey, J.B. (1999). Methods for defining and determining the clinical significance of treatment effects: Description, application, and alternatives. *Journal of Consulting and Clinical Psychology, 67,* 300–307.

Kraus, D.R., Castonguay, L.G. & Wolfe, A. (2006). The Outcomes Assistant: A kinder philosophy to the management of outcomes. *Psychotherapy Bulletin, 41,* 23–31.

Kraus, D.R., Seligman, D. & Jordan, J.R. (2005). Validation of a behavioral health treatment outcome and assessment tool designed for naturalistic settings: The Treatment Outcome Package. *Journal of Clinical Psychology, 61,* 285–314.

Lucas, R.W. (1977). A study of patients' attitudes to computer interrogation. *International Journal of Man-Machine Studies, 9*, 69–86.

MacCallum, R.C., Browne, M.W. & Sugawara, H.M. (1996). Power analysis and determination of sample size for 575 covariance structure modeling. *Psychological Methods, 1*, 130–149.

McHorney, C.A., Ware, J.J. & Raczek, A.E. (1993). The MOS 36-Item Short-Form Health Survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. *Medical Care, 31*, 247–263.

Nagelkerke, N.J.D. (1991). A note on the general definition of the coefficient of determination. *Biometrika, 78*, 691–692.

Nelson, D.C., Hartman, E., Ojemann, P.G. & Wilcox, M. (1995). Breaking new ground: Public/private collaboration to measure and manage Medicaid patient outcomes. *Behavioral Healthcare Tomorrow, 4*, 31–39.

Ogles, B.M., Melendez, G., Davis, D.C. & Lunnen, K.M. (2000). *The Ohio Youth Problem, Functioning, and Satisfaction Scales: Technical Manual.* Columbus, OH: Ohio University.

Searles, J.S., Perrine, M.W., Mundt, J.C. & Helzer, J.E. (1995). Self-report of drinking using touch-tone telephone: Extending the limits of reliable daily contact. *Journal of Studies on Alcohol, 56*, 375–382.

Tabachnick, B.G. & Fidell, L.S. (1996). *Using multivariate statistics* (3rd edn.). New York: Harper Collins College Publishers.

Turner, C.F., Ku L., Rogers, S.M., Lindberg, L.D., Pleck, J.H. et al. (1998). Adolescent sexual behavior, drug use, and violence: Increased reporting with computer survey technology. *Science, 280*, 867–873.

8

Clinical Outcomes in Routine Evaluation (CORE) – The CORE Measures and System: Measuring, Monitoring and Managing Quality Evaluation in the Psychological Therapies

Michael Barkham[1], John Mellor-Clark[2], Janice Connell[3], Chris Evans[4], Richard Evans[5] and Frank Margison[6]

[1]Centre for Psychological Services Research, University of Sheffield, UK, [2]CORE IMS, Rugby, UK, [3]University of Sheffield, UK, [4]University of Nottingham, UK, [5]CORE System Trust, Bath, UK, [6]Manchester Mental Health and Social Care Trust, UK

Introduction

The CORE (Clinical Outcomes in Routine Evaluation) measures make up a battery of client-completed outcome measures derived from a 34-item parent measure – the CORE-OM – which taps the domains of subjective well-being, problems, functioning and risk. The measures share the common method fundamental to all outcome measures identified as *patient reported outcome measures* (PROMs) and which are a central plank in healthcare evaluation. The CORE measures lie at the heart of the broader CORE System that comprises practitioner-completed forms capturing information relating to pre-therapy variables, treatment delivery and post-therapy impacts. This CORE System, in its paper form, is supported by optional software support systems using personal

Developing and Delivering Practice-Based Evidence By Michael Barkham, Gillian E Hardy, and John Mellor-Clark