Validation of a Behavioral Health Treatment Outcome and Assessment Tool Designed for Naturalistic Settings: The Treatment Outcome Package

▼

David R. Kraus and David A. Seligman
Behavioral Health Laboratories

▼

John R. Jordan
Family Loss Project

In 1994, the American Psychological Association and the Society for Psychotherapy Research convened a Core Battery Conference to develop a set of criteria for the selection of a universal core battery that could be used as a common outcome tool across all outcome studies. The Treatment Outcome Package (TOP) is a behavioral health assessment and outcome battery with modules for assessing a wide array of behavioral health symptoms and functioning, demographics, case-mix, and treatment satisfaction. It was developed to follow the design specifications set forth by the Core Battery Conference, but also to ensure the battery's applicability to naturalistic treatment settings in which randomization may be impossible. In this article we discuss a number of studies that evaluate the initial psychometrics of the items that comprise the mental health symptom and functional modules of the TOP. We conclude that the TOP has an excellent factor structure, good test-retest reliability, promising initial convergent and discriminant validity, measures the full range of pathology on each scale, and has some ability to distinguish between behavioral health clients and members of the general population. © 2004 Wiley Periodicals, Inc. J Clin Psychol 61: 285–314, 2005.

Keywords: treatment outcomes; assessment tests; test validation

Editor's Note: This article may represent a conflict of interest. Please see the comment by Karla Moras.

The authors would like to thank Peter Arnett, David Barlow, Tim Brown, Tom Borkovec, two anonymous reviewers, and JoAnn Bouzan for their contributions and suggestions.

Correspondence concerning this article should be addressed to: David R. Kraus, 50 Main Street, Ashland, MA 01721; e-mail: dkraus@bhealthlabs.com.

JOURNAL OF CLINICAL PSYCHOLOGY, Vol. 61(3), 285–314 (2005) © 2005 Wiley Periodicals, Inc.
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/jclp.20084

Mandates for accountability data from behavioral healthcare purchasers and accrediting bodies have led to an industry-wide surge in outcome evaluations in naturalistic, real-world settings. This rapid growth of outcome measurement presents a tremendous opportunity to conduct rigorous, low-cost experimental research using large samples and to improve treatment quality through accurate measurement and appropriate feedback. However, many of these opportunities are predicated on the existence of a core outcome battery that meets the needs of both clinicians and researchers (Borkovec, Echemendia, Ragusea, & Ruiz, 2001). The most noteworthy effort to develop the standards and selection criteria for a core battery comes from the 1994 Core Battery Conference (referred to as Conference) organized by the Society for Psychotherapy Research and the American Psychological Association (Horowitz, Lambert, & Strupp, 1997). The Conference concluded that a Universal Core Battery (UCB) that would work across all diagnostic categories and levels of care was required, and that more focused, diagnostic-specific batteries should supplement the UCB as needed in individual studies. Table 1 summarizes the UCB development and selection criteria.

The Treatment Outcome Package (TOP) was designed to measure outcomes in naturalistic settings and developed using these Conference UCB guidelines. Our purpose here is to present a number of studies that evaluate the TOP in light of the psychometric requirements set forth by the Conference.

Beyond the Conference criteria summarized in Table 1, several other criteria are important in judging the appropriateness of an outcome tool because they affect the chances of widespread adoption in naturalistic settings: the inclusion of case-mix (also known as moderator or risk-adjustment) variables (Goldfield, 1999), minimal or absent floor and ceiling effects, and real-time reporting. Our rationale for including each of them is discussed below.

Variables that are beyond the control of the therapeutic process, but nonetheless influence the outcome, are defined as case-mix variables (Goldfield, 1999). Naturalistic

Table 1
Universal Core Battery Requirements

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 nonpatients
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 et al., 1997).

research typically lacks the experimental controls used in efficacy research to mitigate against the need to statistically control for (or measure) these case-mix variables.[1] 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. Furthermore, a major, and essential purpose of naturalistic outcome research is to assess the generalizability of tightly controlled efficacy research to real-world settings. In order to discover the populations to which the efficacy results can generalize, case-mix must obviously be measured.

If an outcome tool is to be truly applicable across all diagnostic groups and consumer populations, it needs to demonstrate that it can measure the full spectrum of pathology. Since this is rarely discussed in convergent validation samples (cf. Foa, Kozak, Salkovskis, Coles, & Amir, 1998), we believe analysis of floor and ceiling effects should be a separate psychometric requirement. 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. The use of measures that cannot do this is comparable to the use of a basal body thermometer (with a built-in ceiling of only 102°) 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°. 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 significant changes can lead to poor administrative and clinical decisions. Floor and ceiling effects of the TOP are discussed in Study 4 of this article.

Although not part of an outcome tool per se, near-real-time delivery of results to clinicians is imperative. The Conference hints at this by noting that the tool and its results should be clinically useful. 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. For both the client and the clinician, the purpose of participation in naturalistic outcome studies must be first and foremost to help the client to improve and only secondarily to assist a research project. Therefore, an outcome tool, its

[1] In randomized controlled trials, risk adjustment is typically neither necessary nor done; yet in naturalistic outcomes it is essential. In randomized controlled trials, certain case-mix variables are seen as so powerful that they are typically controlled by making them part of strict inclusion and exclusion criteria (e.g. co-morbid medical conditions). Strict inclusion and exclusion criteria in efficacy research limit the variance in important case-mix variables. Further mitigating the effects of case-mix variables, random assignment increases the chances that other uncontrolled variables are evenly distributed across control conditions.

However, most naturalistic research studies lack one or both of the methods that tend to homogenize the samples (random assignment and strict inclusion exclusion criteria). The samples in naturalistic outcome measurement are often quite heterogeneous and require larger *N*s, disaggregation, and/or statistical techniques to control for sample differences. Consequently, measurement tools like the Beck Depression Inventory or SCL-90 that have been used successfully for decades in efficacy research are simply insufficient (by themselves) for naturalistic outcome measurement because these tools lack case-mix variables, making risk-adjustment or disaggregation impossible. Measurement of symptoms, functioning, and general distress must be augmented by extensive assessment of case-mix.

processing system, and report structures must deliver useful and immediate feedback. The mean time it takes for a report to be returned to the provider after a TOP is completed is 16 minutes.

The Treatment Outcome Package

Since a battery meeting all of the above criteria did not exist, a decision was made to develop a new battery. Additional rationale for creating a new instrument included the demand for a royalty-free set of modules, creation of one common Likert scale for all clinical questions, elimination of duplicate items across measures, and control over the rights to modify and add questions in the future.

Initial development of these modules consisted of the first author generating more than 250 a-theoretical items that spanned diagnostic symptoms and functional areas identified in the *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition* (*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 their review and editing. 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. The instrument presented here is the result of four iterations of that process.

The current version of the TOP is a battery of distinct modules that can be administered all together or in combinations as needed. Expert clinical and client review in the development process ensured adequate face validity. The various modules of the TOP include:

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

The present study focuses on the 93 items of the TOP designed to measure functioning, quality of life, and mental health symptoms since these scales are directly related to the Conference criteria for a UCB. In this article six separate studies are included, each impacting upon UCB criteria. Study 1 determines the factor structure of the TOP to ensure a theory free tool with a solid foundation derived from extensive patient populations spanning all levels of care. Study 2 provides preliminary information on the test-retest reliability of the TOP scales determined in Study 1. Study 3 explores the discriminant and convergent validity, comparing the TOP to other standard assessment and outcome tools in the industry. Study 4 explores floor and ceiling effects of the TOP that are critical factors in ensuring the scales applicability to diverse clinical populations and its sensitivity to change. Study 5 specifically explores the scale's sensitivity to change, while Study 6 explores the tool's ability to distinguish patients from nonpatients.

Instruments

The following instruments were used in the present studies to test the validity of the TOP.

The Beck Depression Inventory

The Beck Depression Inventory (BDI; Beck, Steer, & Ranieri, 1988) is a 21-item self-report scale used to assess cognitive and physical symptoms of depression. It has been used extensively in psychological research with numerous populations and psychiatric disorders. The BDI has been a central outcome tool used in depression efficacy research and was specifically recommended as a good example of the measurement of depression by the Core Battery Conference. Mean internal consistency across patient and nonpatient samples is .86 (Beck, Steer, & Garbin, 1988), and there is good validation data with other measures of depression like the Hamilton Psychiatric Rating Scale for Depression (.73), the Zung Self-Reported Depression Scale (.76), and the MMPI Depression Scale (.76) (Groth-Marnat, 1990).

The Brief Symptom Inventory

The Brief Symptom Inventory (BSI; Derogatis, 1975) is a 53-item version of the Symptom Checklist-90-Revised (SCL-90-R; Derogatis, 1977). The SCL-90-R succeeded the Hopkins Symptom Checklist (HSCL), which was used as part of the core outcome battery developed in 1970 by the National Institute for Mental Health (Waskow, 1975). The BSI is a self-report scale that has been used to assess a broad array of psychiatric symptoms. It has been used extensively in psychological research across numerous populations and psychiatric disorders (Flynn, 2002; Trabin, Freeman, & Pallak, 1995). Its scales include Somatization, Obsessive–Compulsive, Interpersonal Sensitivity, Depression, Anxiety, Hostility, Phobic Anxiety, Paranoid Ideation, and Psychoticism. The internal consistency of BSI scales ranges from .71 (Psychoticism) to .85 (Depression). Except for somatization (.68), test-retest reliabilities are good (.78 to .85) (Derogatis, 1975). The BSI's validity has been extensively tested against the SCL-90-R, MMPI (Hathaway & McKinley, 1989), and many others.

The Minnesota Multiphasic Personality Inventory-2

The MMPI-2 (Hathaway & McKinley, 1989) is a 567-item self-report instrument used to assess personality characteristics. The original MMPI was also part of the core outcome battery developed by NIMH in 1970 (Waskow, 1975). The MMPI-2 has been extensively used in psychological research across numerous populations and psychiatric disorders. Reliability of MMPI-2 scales is mixed with certain scales (5, 6, 9) showing unacceptable levels of internal consistency, which is further supported by the lack of unidimensional scales in factor analysis. Other scales (especially 1, 7, 8, 0) have good internal consistency and will be emphasized in the analyses below (Graham, 1993).

The BASIS 32

The BASIS 32 (Eisen, Grob, & Klein, 1986) is a 32-item self-report instrument designed to measure clinical outcomes in inpatient facilities. As evidenced by its inclusion by most outcome software vendors, it has emerged as one of the most widely used naturalistic outcome tools (Trabin et al., 1995), and studies have documented its utility in outpatient settings as well (Eisen, Wilcox, Leff, Schaefer, & Culhane, 1999). The BASIS 32 has five scales (Depression and Anxiety, Relation to Self and Others, Psychosis, Impulsive and Addictive Behavior, Daily Living and Role Functioning) and one summary scale (Overall Score). Test-retest reliability of the BASIS-32 total score was .85 with specific subscales' reliabilities for Relation to Self and Others at .80; Daily Living Skills and Role Functioning at .81; Depression and Anxiety at .78; Impulsive and Addictive Behavior at .65; and Psychosis at .76 (Eisen, Dill, & Grob, 1994). Correlating scores with hospitalization status and the tool's ability to discriminate between diagnostic groups have assessed the tool's validity.

The SF-36

The SF-36 (Ware & Sherboume, 1992) is a 36-item self-report measure of general health status. It has eight scales and two summary scales (Mental and Physical). As evidenced by its inclusion by most outcome software vendors, the SF-36 has also emerged as one of the most widely used psychiatric outcome scales in naturalistic settings. It has documented satisfactory reliability and validity in both psychiatric and medical settings. Internal consistency of SF-36 subscales is generally reported to exceed .80 (Jette & Downing, 1994; Garratt, Ruta, Abdalla, Buckingham & Russell, 1993; Ware, 1996). Its usefulness as a general outcome tool has been tested on a wide variety of mental health and general medical patients (e.g. Brazier et al., 1992; McHorney, Ware, & Raczek, 1993; Wells et al., 1989).

Study 1: Factor Structure

In this section, we describe the exploratory factor analysis (EFA) and confirmatory factor analyses (CFA) of the TOP's internal structure.

Method

For this study, 93 mental health symptom, functional, and quality of life items administered to a large sample of newly admitted psychiatric clients were analyzed. Participants were instructed to rate each question in relation to "How much of the time during the last month you have a . . ." 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 behavioral health services across the United States who completed all questions of the TOP at intake as part of standard treatment. Age, sex, and years of education for the samples are summarized in Table 2. Breakdown of service facility types is presented in Table 3.

Procedure

The sample was split into five random subsamples (split sample 1, 2, 3, 5, *n*'s = 3,960 and sample 4, *n* = 3,961) as a cross-validation strategy. Samples 1 and 2 were used to develop

Table 2
Participants

	M	*SD*
Study 1: Factor Analysis		
19,801 Participants		
Age	33.3	12.1
Education	12.4	3.7
% Women	51%	
Study 2: Test-Retest		
53 Participants		
Age	38.6	15.4
Education	11.1	6.2
% Women	63%	
Study 3: Discriminant and Convergent Validity		
312 Participants		
Age	41.2	16.2
Education	14.0	3.6
% Women	65%	
Study 4: Floor and Ceiling Effects		
216,642 Participants		
Age	33.8	13.8
Education	12.0	4.1
% Women	58%	
Study 5: Sensitivity to Change		
20,098 Participants		
Age	33.1	14.2
Education	12.1	4.3
% Women	60%	
Study 6: Criterion Validity		
1,034 Participants		
Age	46.3	17.5
Education	14.9	3.4
% Women	74%	

Table 3
Psychiatric Services Included in Studies 1, 4, and 5

	Number of Facilities in Each Study		
Service Type	Study 1	Study 4	Study 5
Long-term inpatient locked units	3	10	4
Acute short-term inpatient locked units	21	39	26
Acute short-term inpatient unlocked units	11	16	12
Partial hospitalization programs	20	70	51
Crisis stabilization/respite programs	10	13	8
Crisis/emergency evaluation	27	34	26
Outpatient milieu programs (e.g., day treatment)	21	61	27
Outpatient therapy programs	183	379	206
Outpatient assessment and referral services	5	24	5
Community living/supported housing	12	32	14
Residential programs	55	130	84
Employee assistance programs	2	2	2
Unknown	13	115	46

and refine a factor model that was subsequently confirmed in samples 3–5. All samples had no missing data.

Sample 1 was used to develop a baseline factor model. Responses to the 93 items (detailed in Table 4) 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 1 supplemented by the scree test and the criterion of interpretability (Cattell, 1966; Tabachnick & Fidell, 1996). Items that did not load on at least one factor greater than 0.45, 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). 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.

Results

Based on the above criteria, 16 factors were extracted and reviewed from the EFA dataset. Both orthogonal and oblique rotations were explored, with only minor differences found in factor loadings. With the assumption that these factors are related to each other, the oblimin rotation was chosen. Five factors were dropped due to insufficient number of loading items. Other items were also trimmed due to insufficient loadings. In total, 41 of the 93 items were dropped and the final 52 items were again analyzed with oblique rotations. The final model produced 11 factors accounting for 63% of the variance and is presented in Table 5.

This model was then tested using structural equation modeling in an initial CFA (Sample 2). Although this model did meet the conservative multiple-index fit criteria (Table 6), fit diagnostics indicated that the model could be improved. Through exploring all possible sources of strain (potential cross loadings, method effects, over- or under-factoring, and minor factors), a series of steps were taken to improve the model, now using the CFA framework in an exploratory fashion. With each modification, the χ^2_{diff} was significant ($p < 0.001$). During this process, four additional items (1, 4, 29, 55) were eliminated from the model due to relatively low factor loadings and the item having more than one correlated error with items on other factors. In these cases, dropping the item from the model improved overall model fit. The model was also improved by freeing ten items to crossload on other factors. Standardized regression weights for these cross-loaded items ranged from 0.132 to 0.315 with a mean of 0.200. The crossloading items can be seen in Table 4. Finally, eight correlated errors were mapped into the model. Two were mapped due to item juxtaposition (90–91, 47–48), four were mapped due to item content similarity (68–70, 64–66, 56–60, 23–26), and two for both reasons (24–25,

Table 4
Original 93 TOP Items With Final Primary and Secondary Factor Loading

Item	Item Wording	Final Model
1	Been satisfied with your physical abilities	
2	Felt a lack of closeness or contact with others	
3	Been satisfied with your relationships with others	LIFEQ
4	Been satisfied with your sleep	
5	Been satisfied with your daily responsibilities	LIFEQ
6	Been satisfied with your sex life	
7	Been satisfied with your general mood and feelings	LIFEQ
8	Been satisfied with how you cope with daily problems	
9	Been satisfied with your life in general	LIFEQ
10	Had trouble telling others your feelings or needs	
11	Felt that others were not responding to your feelings or needs	
12	Felt too much conflict with someone	SCONF
13	Been emotionally hurt by someone	SCONF
14	Felt someone else had too much control over your life	SCONF
15	Felt too dependent on others	
16	Had trouble falling asleep	SLEEP
17	Had nightmares	SLEEP, PSYCS
18	Awakened frequently during the night	SLEEP
19	Had trouble returning to sleep after awakening in the night	SLEEP
20	Felt tired during the day	
21	Slept too much or at unwanted times	
22	Had conflicts with others at work or school regardless of fault	WORKF
23	Missed work or school for any reason	WORKF
24	Not been acknowledged for your accomplishments	WORKF
25	Had your performance criticized	WORKF
26	Not been excited about your work or school work	WORKF
27	Spent too much time working	
28	Yelled at someone	
29	Broken or damaged things in anger	
30	Physically hurt someone else or an animal	VIOLN
31	Had desires to seriously hurt someone	VIOLN
32	Had thoughts of killing someone else	VIOLN
33	Felt that you were going to act on violent thoughts	VIOLN, SUICD
34	Felt no desire for, or pleasure in, sex	SEXFN
35	Had sexual thoughts you did not want to have	
36	Felt sexually incompatible with your partner or frustrated by the lack of a partner	SEXFN, SCONF
37	Felt emotional or physical pain during sex	SEXFN

(*continued*)

Table 4
Continued

Item	Item Wording	Final Model
38	Been aroused by things that felt unacceptable	▨
39	Had trouble functioning sexually (having orgasms, etc.)	SEXFN
40	Felt shaky or trembled	▨
41	Had a racing heart	PANIC
42	Felt light-headed	PANIC
43	Frequently urinated	▨
44	Had shortness of breath	PANIC
45	Been startled (by a touch or by someone entering the room)	▨
46	Felt nauseous, had diarrhea or other stomach or abdominal pains	▨
47	Had a dry mouth or trouble swallowing (“a lump in your throat”)	PANIC
48	Had sweaty hands (clammy) or cold hands or feet	PANIC
49	Felt restless, keyed up, or on edge	▨
50	Had muscle pain, including back, neck, or headache pain	▨
51	Felt down or depressed	DEPRS
52	Felt easily irritated or annoyed	▨
53	Felt little or no interest in most things	DEPRS
54	Felt hopeless	▨
55	Felt nervous or anxious	▨
56	Felt guilty	DEPRS
57	Felt angry	▨
58	Felt restless	DEPRS, MANIA
59	Wanted to be alone	▨
60	Felt worthless	DEPRS, SUICD
61	Had to do something to avoid anxiety or fear (washing hands, etc.)	▨
62	Felt shy or inhibited	▨
63	Felt tired, slowed down, or had little energy	DEPRS
64	Worried about things	DEPRS
65	Had trouble concentrating or making decisions	DEPRS
66	Noticed your thoughts racing ahead	DEPRS, MANIA
67	Been too talkative	▨
68	Inflicted pain on yourself	SUICD
69	Felt rested after only a few hours of sleep	MANIA, PSYCS
70	Thought about killing yourself or wished you were dead	SUICD, DEPRS
71	Planned or tried to kill yourself	SUICD
72	Avoided certain situations due to fear or panic	▨
73	Felt emotionally numb to something that would normally cause intense feelings	▨
74	Felt you were better than other people	MANIA, WORKF

(*continued*)

Table 4
Continued

Item	Item Wording	Final Model
75	Felt on top of the world	MANIA
76	Felt panic in places that would be hard to leave if necessary	▨
77	Had a large appetite or little or no appetite	▨
78	Had trouble with your memory	▨
79	Felt others were working against you	▨
80	Had no time for yourself	▨
81	Felt responsible for your troubles	▨
82	Worried that someone might hurt you	PSYCS, SCONF
83	Felt detached from what was really happening	▨
84	Been unable to talk to at least one other person about your problems	▨
85	Had unwanted thoughts or images	PSYCS
86	Worried about going crazy	▨
87	Done something without thinking of the consequences	▨
88	Felt people or events kept you from achieving your goals	▨
89	Felt confused, in a fog, or dazed	▨
90	Seen or heard something that was not really there	PSYCS
91	Felt someone or something was controlling your mind	PSYCS
92	Forced yourself to throw-up food	▨
93	Had difficulty remembering personal information (important life events or periods of time)	▨

Note. ▨ Dropped items.

65–66). All modifications to the model were made based on both strain indices and the conceptual interpretation of the findings.

Samples 3, 4, and 5 were used to validate the final model developed with Sample 2, and showed excellent and consistent model fit criteria across all indices. Taken together, there is strong support for the stability and strength of these factors. Results are summarized in Table 6 and demonstrate excellent model fit with no significant strains. The factor names, Cronbach's alphas to assess internal consistency, and intercorrelations are listed in Table 7.

Discussion

The TOP was designed to assess a broad range of behavioral health functional and symptom domains. The factor analysis presented here revealed eleven stable and clinically useful TOP subscales with excellent confirmatory modeling in large samples of diverse patients. One factor (Quality of Life) incorporates questions about how often the client has felt satisfied with various areas of his or her life (e.g. "been satisfied with your life in general"). Three other factors include functional questions and are labeled: Work Functioning, Sexual Functioning, and Social Conflict. The other seven factors hold symptom

Table 5
EFA Pattern Matrix

	Component										
Item	DEPRS	VIOLN	WORKF	LIFEQ	SLEEP	SEXFN	SCONF	SUICD	MANIA	PSYCS	PANIC
64	.649										
56	.622										
55	.618										−.264
66	.595										
65	.588										
58	.578										
51	.577										
53	.556										
60	.535										
63	.455										
31		.817									
32		.769									
33		.757									
30		.731									
29		.606									
26			.699								
22			.660								
25			.651								
23			.649								
24			.645								
3				.747							
5				.736							
9				.681							
1				.644							
7	−.257			.635							
19					−.864						
18					−.859						
16					−.798						
4				.531	.566						
17					−.493					−.274	
39						.776					
34						.689					
37						.677					
36						.666					
13							−.746				
12							−.720				
14							−.706				
71								.904			
68								.740			
70								.738			
74									.761		
75									.745		
69									.453		
91										−.746	
90										−.691	
82										−.609	
85	.322									−.503	
44											−.770
42											−.697
47											−.677
41											−.667
48											−.627

Note. Only loadings greater than 0.25 are shown. Dropped items.

Table 6
CFA Validation

CFA	Description	*N*	DF	TLI	CFI	RMSEA	RMSEA Upper
Sample 2 initial	Derived from EFA model	3,960	1218	.898	.906	.045	.046
Sample 2 final	Modified model	3,960	1007	.945	.951	.033	.034
Sample 3	Confirmatory analysis 1	3,960	1007	.940	.946	.035	.036
Sample 4	Confirmatory analysis 2	3,960	1007	.942	.948	.034	.035
Sample 5	Confirmatory analysis 3	3,961	1007	.940	.947	.035	.036

items and include: Depression, Panic, Psychosis, Suicidal Ideation, Violence, Mania, and Sleep. Cronbach's alphas were used as one estimate of scale reliability and were adequate for all factors, with the exception of Mania. Its lower internal consistency may be due to the nature of the items that load onto it, in which extreme scores at either end may be viewed as unhealthy (symptoms of mania or depression), while scores in the middle may be viewed as healthy. Despite its questionable internal consistency, Mania was retained as a factor because of its clinical importance and acceptable test-retest reliability (see Study 2 below).

Through this method of developing the factor structure of the TOP, it is clear that many clinically interesting questions have been dropped (as compared to previous versions of the tool). If it becomes clear from additional research and clinician feedback that these questions are valuable, it may be important to return these items with additional questions from the same construct and develop additional factors for inclusion in future versions. In other words, these questions may have been related to important clinical constructs for which insufficient items were available to form reliable factor structures.

While the TOP includes many questions about both the physiological and cognitive components of anxiety, just the physiological symptoms of anxiety loaded on a separate factor, which we labeled *Panic*. Some cognitive symptoms of anxiety (e.g., worried about things, noticed your thoughts racing ahead, felt restless) loaded on the Depression factor, a finding consistent with the literature (Barlow, Bach, & Tracey, 1998; Eisen, Grob, & Klein, 1986).

Finally, space limitations prevent an adequate review of factor invariance analyses of the TOP factors in this large clinical population. Analyzing whether people in different demographic and clinical populations show similar patterns of responses is an important discussion to which an entire article could be devoted.

Study 2: Test-Retest Reliability

In this section, we report on the test-retest reliability of the TOP. Another measure of reliability, internal consistency, was presented in Study 1.

Method

In 1998, 53 behavioral health clients were recruited by four community mental health centers to participate in a one-week test re-test study. All clients were Medicaid enrollees who completed the Treatment Outcome Package one week apart while they were waiting

Table 7
Factor Correlation Matrix and Test-Retest Reliabilities

Factor	Description	DEPRS	VIOLN	SCONF	LIFEQ	SLEEP	SEXFN	WORKF	PSYCS	PANIC	MANIC	SUICD	α	Intraclass Test-Retest
DEPRS	Depression	1.00											.93	.93
VIOLN	Violence	0.33	1.00										.81	.88
SCONF	Social Conflict	0.55	0.33	1.00									.72	.93
LIFEQ	Quality of Life	−0.78	−0.24	−0.45	1.00								.85	.93
SLEEP	Sleep Functioning	0.64	0.26	0.41	−0.50	1.00							.86	.94
SEXFN	Sexual Functioning	0.51	0.21	0.38	−0.41	0.36	1.00						.69	.92
WORKF	Work Functioning	0.55	0.43	0.53	−0.41	0.37	0.34	1.00					.72	.90
PSYCS	Psychosis	0.66	0.55	0.42	−0.46	0.51	0.42	0.50	1.00				.69	.87
PANIC	Panic	0.73	0.33	0.43	−0.52	0.59	0.43	0.46	0.67	1.00			.83	.88
MANIC	Mania	−0.26	0.11	−0.09	0.37	−0.12	−0.09	0.01	0.05	0.04	1.00		.53	.76
SUICD	Suicidality	0.44	0.44	0.26	−0.33	0.27	0.23	0.36	0.61	0.38	−0.02	1.00	.78	.90

for outpatient treatment to begin. Age, sex, and years of education for the sample are summarized in Table 2.

Results

The stability of the TOP over time was assessed by computing intraclass correlation coefficients using a one-way random model. Except for MANIA, all reliabilities for subscales (factors presented in Study 1) were excellent (see Table 7), ranging from .87 to .94. Mania's reliability was acceptable, but considerably lower at .76.

Discussion

Results of Study 2 revealed that the TOP has good test-retest reliability for all subscale scores for the sample chosen. However, the sample used (53 outpatients waiting for treatment to begin), was chosen largely because it was easy to obtain. Ideally, it would be important to assess test-retest reliabilities for all types of populations and levels of care. However, for some levels of care with acute or high-risk clients, it is difficult or impossible to ethically obtain a sample waiting for treatment. One potential source of participants representing a more severe population that might ethically be recruited would be a homeless population with previous severe psychiatric histories who are currently refusing treatment. Until more diverse and larger samples are collected, all that may be stated is that the TOP seems to have good test-retest reliability among outpatients awaiting clinical treatment. The use of intraclass correlation in this analysis demonstrates that not only is the rank-ordering of clinical severity in patients similar from one week to the next, but so is the actual level of severity within patients.

Study 3: Discriminant and Convergent Validity

In this section, we evaluate the discriminant and convergent validity of the factors developed in Study 1. Important to the testing of the validity of a measure is the testing of whether the measure 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). The validity instruments chosen for this study were selected because of their acceptable psychometrics and prominence in the field. Because the instruments were chosen before the factors from Study 1 emerged, a few factors do not have ideal convergent validity measures. In evaluating the results, it should be noted that there is no item overlap between the TOP and any of the validity measures—if such overlap did exist, it might artificially inflate the convergent correlations.

Method

Study 3 included 312 participants. Age, sex, and years of education for the sample are summarized in Table 2. Ninety-four participants were from the general population, 123 were from an outpatient clinical population, and 95 were from an inpatient clinical population. All 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). Ideally, all patients would have completed all measures, however attempting this may have represented too large a burden for many participants. That all patients did not complete all measures should be considered in interpreting the results. Specifically, it suggests that differences between validity scales in the relative magnitude of correlations may be due to sample differences (or tool reliability differences) rather than to differences in true relationships among the constructs.

All clients signed informed consent and were recruited through customers of BHL. During a 2-month period in 1996, the first author attempted to recruit all newly admitted patients within the first 24 hours of admission to three inpatient psychiatric and substance abuse units in a Boston area hospital. Outpatient clinicians who agreed to participate in the study attempted to recruit all new admissions during a 6-month period during 1997. General population samples were recruited by clinicians from BHL sites during 1997 by asking friends and acquaintances (nonfamily) to participate in the study.

Procedure

Validity scales were used to evaluate discriminant and convergent validity of the TOP. The specific measures used for both are detailed in the results section below. Because some of the TOP factors are not normally distributed, both Pearson and Spearman correlations were analyzed and reviewed. No significant differences were found between the two methods and just the Pearson correlations are presented below.

Results

The construct validity of the TOP was assessed by correlating each TOP measure with each validity measure's score. The entire correlation matrix is presented in Table 8. Because the study design did not call for all clients to complete all measures, it is impossible to evaluate the relative strength of correlations between some of the validity measures and the TOP. Therefore, these differences are not discussed.

As discussed below, inspection of the correlations indicated that the TOP measures generally showed the expected relationships with other relevant self-report measures of psychiatric symptoms and functioning. In most cases, convergent coefficients were quite high for each validity measure.

Measuring depression, the TOP Depression (DEPRS) scale should show convergent relationships with other measures of the same construct. These are: the BDI (.92), MMPI-Depression (.73), BSI-Depression (.90), BSI-Anxiety (.70), BASIS32-Depression/Anxiety (.86), the SF36-Mental Health (.82), and the SF-36-Vitality (.68) measures. All of these correlations were quite high. By contrast the TOP Depression scale should not correlate with MMPI-Mania (−.23), or the MMPI-Schizophrenia (.24) measures.

Measuring violence and temper, the TOP Violence (VIOLN) scale was expected to correlate with the BSI-Hostility scale (.77). A similar, but not identical construct is tapped by the BASIS32-Impulsive (.69). It was not expected to correlate with MMPI-Hypochondriasis (−.16) or BSI Somatization (−.13).

Measuring interpersonal functioning and conflict, the TOP Social Conflict (SCONF) scale was expected to correlate with the BASIS32-Relationship to Self and Other (.60), SF36-Social Functioning (−.35), MMPI Social Introversion (.37), BSI-Paranoid (.72), and BSI-Interpersonal Sensitivity (.44). It was not expected to correlate with BSI-OCD (−.24), or MMPI-Psychasthenia (−.04).

Measuring quality of life and subjective distress, the TOP Quality of Life (LIFEQ) scale was expected to correlate with SF36-Vitality (−.57), SF36-Mental Health (−.68),

Table 8
Correlations Between TOP and Validity Scales

	DEPRS	VIOLN	SCONF	LIFEQ	SLEEP	SEXFN	WORKF	PSYCS	PANIC	MANIC	SUICD
BDI	−.92** (66)	−.41** (66)	−.62** (65)	.71** (63)	−.52** (67)	−.42** (57)	−.37** (63)	−.50** (67)	−.49** (65)	−.08 (65)	−.60** (65)
MMPI											
HS	−.09 (66)	−.27* (66)	−.14 (65)	.19 (63)	−.11 (67)	.00 (57)	−.34** (63)	−.02 (67)	−.30* (65)	−.01 (65)	−.05 (65)
D	−.73** (66)	−.27* (66)	−.46** (65)	.60** (63)	−.43** (67)	−.31* (57)	−.25* (63)	−.47** (67)	−.30* (65)	−.11 (65)	−.42** (65)
HY	−.29* (66)	−.16 (66)	−.25* (65)	.29* (63)	−.42** (67)	−.18 (57)	−.21 (63)	−.17 (67)	−.13 (65)	−.21 (65)	−.15 (65)
PD	−.41** (66)	−.05 (66)	−.38** (65)	.59** (63)	−.03 (67)	−.33* (57)	−.16 (63)	−.22 (67)	−.17 (65)	.19 (65)	−.27* (65)
PA	−.51** (66)	−.19 (66)	−.38** (65)	.43** (63)	−.14 (67)	−.37** (57)	−.24 (63)	−.36** (67)	−.32** (65)	−.08 (65)	−.37** (65)
PT	.03 (66)	.14 (66)	.00 (65)	.05 (63)	−.01 (67)	−.14 (57)	.11 (63)	.00 (67)	.20 (65)	.04 (65)	−.04 (65)
SC	−.24 (66)	−.05 (66)	−.28* (65)	.28* (63)	−.20 (67)	−.27* (57)	−.27* (63)	−.28* (67)	−.13 (65)	−.06 (65)	−.22 (65)
MA	.23 (66)	.15 (66)	.13 (65)	−.23 (63)	−.10 (67)	.07 (57)	.03 (63)	.18 (67)	.16 (65)	.43** (65)	.18 (65)
SI	−.28* (64)	−.20 (65)	−.37** (64)	.30* (63)	−.17 (65)	−.13 (56)	−.26* (62)	−.24 (66)	−.12 (64)	−.14 (64)	−.08 (64)
BSI											
Depression	−.90** (66)	−.42** (66)	−.61** (65)	.66** (63)	−.46** (67)	−.36** (57)	−.37** (63)	−.50** (67)	−.50** (65)	−.16 (65)	−.69** (65)
Psychoticism	−.63** (66)	−.27* (66)	−.53** (65)	.50** (63)	−.36** (67)	−.24 (57)	−.46** (63)	−.72** (67)	−.60** (65)	−.18 (65)	−.46** (65)
Somatization	−.30* (66)	−.13 (66)	−.47** (65)	.43** (63)	−.22 (67)	−.29* (57)	−.32** (63)	−.39** (67)	−.72** (65)	.02 (65)	−.30* (65)
Hostility	−.41** (66)	−.77** (66)	−.41** (65)	.17 (63)	−.34** (67)	−.12 (57)	−.32** (63)	−.37** (67)	−.18 (65)	−.18 (65)	−.51** (65)
Phobic	−.52** (66)	−.25* (66)	−.52** (65)	.45** (63)	−.27* (67)	−.21 (57)	−.40** (63)	−.56** (67)	−.82** (65)	−.18 (65)	−.35** (65)
OCD	−.29* (66)	−.16 (66)	−.24 (65)	.31* (63)	−.10 (67)	−.10 (57)	−.31* (63)	−.44** (67)	−.19 (65)	.01 (65)	−.21 (65)
Anxiety	−.70** (66)	−.20 (66)	−.42** (65)	.54** (63)	−.38** (67)	−.30* (57)	−.23 (63)	−.36** (67)	−.41** (65)	−.14 (65)	−.47** (65)
Interpersonal	−.38** (66)	−.15 (66)	−.44** (65)	.35** (63)	−.27* (67)	−.21 (57)	−.41** (63)	−.45** (67)	−.60* (65)	−.10 (65)	−.23 (65)
Paranoid	−.58** (66)	−.28* (66)	−.72** (65)	.45** (63)	−.44** (67)	−.35** (57)	−.47** (63)	−.47** (67)	−.42** (65)	−.08 (65)	−.38** (65)
Basis 32											
Rel Self Other	−.84** (110)	−.58** (110)	−.60** (36)	.71** (110)	−.54** (110)	−.28** (105)	−.56** (102)	−.61** (110)	−.68** (110)	−.15 (109)	−.64** (110)
Daily Role	−.82** (110)	−.57** (110)	−.65** (36)	.73** (110)	−.46** (110)	−.24* (105)	−.51** (102)	−.63** (110)	−.67** (110)	−.12 (109)	−.67** (110)
Dep/Anx	−.86** (110)	−.61** (110)	−.44** (36)	.73** (110)	−.61** (110)	−.22* (105)	−.55** (102)	−.68** (110)	−.73** (110)	−.22* (109)	−.72** (110)
Impulsive	−.68** (110)	−.69** (110)	−.15 (36)	.57** (110)	−.49** (110)	.00 (105)	−.49** (102)	−.68** (110)	−.62** (110)	−.20* (109)	−.59** (110)
Psychosis	−.62** (110)	−.58** (110)	−.41* (36)	.52** (110)	−.45** (110)	−.21* (105)	−.53** (102)	−.80** (110)	−.59** (110)	−.22* (109)	−.64** (110)
SF-36											
Physical Func	.28* (77)	.33** (76)	.30 (32)	−.26* (77)	.29* (77)	.39** (75)	.11 (70)	.27* (77)	.31** (77)	.33** (76)	.19 (77)
Role Physical	.43** (77)	.31** (76)	.30 (32)	−.42** (77)	.39** (77)	.41** (75)	.38** (70)	.45** (77)	.47** (77)	.25* (76)	.19 (77)
Bodily Pain	.39** (77)	.26* (76)	.37* (32)	−.38** (77)	.46** (77)	.31** (75)	.21 (70)	.36** (77)	.42** (77)	.23* (76)	.23* (77)
General Health	.48** (78)	.24* (77)	.25 (32)	−.56** (78)	.37** (78)	.29* (76)	.31** (71)	.44** (78)	.44** (78)	.27* (77)	.32** (78)
Vitality	.68** (78)	.36** (77)	.56** (32)	−.57** (78)	.39** (78)	.47** (76)	.18 (71)	.45** (78)	.54** (78)	.22 (77)	.51** (78)
Social Func	.75** (78)	.43** (77)	.35 (32)	−.68** (78)	.61** (78)	.37** (76)	.44** (71)	.57** (78)	.50** (78)	.28* (77)	.53** (78)
Role Emotional	.59** (75)	.42** (74)	.22 (30)	−.51** (75)	.49** (75)	.36** (73)	.39** (68)	.59** (75)	.49** (75)	.20 (74)	.36** (75)
Mental Health	.82** (78)	.48** (77)	.54** (32)	−.68** (78)	.47** (78)	.36** (76)	.34** (71)	.54** (78)	.62** (78)	.32** (77)	.69** (78)

*Significant $p < 0.05$ (*N* in parentheses).
**Significant $p < 0.01$ (*N* in parentheses).

and SF36-General Health (−.56). It was not expected to correlate with BSI-OCD (.31), or MMPI-Hypochondriasis (.29).

No validity instruments had a direct measure of the construct of sleep disturbance. However, the TOP Sleep (SLEEP) scale was expected to correlate with other measures that relate to sleep functioning, including SF36-Bodily Pain (.46), SF-36 Vitality (.40), BDI (−.52), BSI-Depression (−.46), MMPI-Depression (−.43), and BASIS32-Depression/Anxiety (−.61). It was not expected to correlate with MMPI-Psychopathic Deviance (−.03), or MMPI-Schizophrenia (−.20).

With no direct validity measure of sexual functioning, the TOP Sexual Functioning (SEXFN) scale was not expected to correlate highly with any validity measure, but it was expected to correlate moderately with several measures that are related to sexual functioning like the BDI (.42), SF36 Vitality (.47), and other measures of depression [MMPI-Depression (−.31), BSI-Depression (.36), and the BASIS32-Depression/Anxiety (−.22)].

Measuring work performance and functioning, the TOP Work Functioning (WORKF) scale was expected to correlate with BASIS32-Daily Role (−.51), and the SF36-Role Functioning Emotional (.40). It was not expected to correlate with MMPI-Psychasthenia (−.11).

Measuring issues related to psychotic processes, the TOP Psychosis (PSYCS) scale was expected to correlate with MMPI-Schizophrenia (−.28), BSI-Psychoticism (.72), and the BASIS32-Psychosis (.80). It was not expected to correlate with MMPI-Hypochondriasis (−.17), or MMPI-Mania (.18).

Measuring the physiological symptoms of anxiety, the TOP Panic (PANIC) scale was expected to correlate with BSI-Somatization (.67), BSI-Anxiety (.50), BASIS32-Depression/Anxiety (.82), and SF36-Vitality (−.65). It was not expected to correlate with MMPI-Psychopathic deviate (.30), or BSI-Hostility (.23).

Measuring symptoms of mania, the TOP Manic (MANIC) scale was expected to correlate with the MMPI-Hypomania (.43) scale, and was not expected to correlate with MMPI-Hypochondriasis (−.21), or the MMPI-Psychasthenia (.04) scales.

Measuring suicidal ideation and planning, the TOP Suicide (SUICD) scale was expected to correlate with related measures of depression like the BDI (.60), BSI-Depression (.69), and BASIS32-Depression/Anxiety (.72). It was not expected to correlate with MMPI-Hypochondriasis (−.15), or the MMPI-Psychasthenia (−.04) scales.

Discussion

This first study evaluating the validity of the TOP provides an initial foundation of data on the TOP factors. Two limitations to the current study that should be addressed in future work are the lack of validity measures that tap directly suicidality, sleep, and sexual functioning, and the failure to have all clients complete all validity measures. However, these limitations do not prevent one from drawing important initial conclusions about the TOP's convergent and discriminant validity.

As an initial study, these results document good convergent and excellent discriminant ability of many of the TOP scales. Indeed, almost all expected convergent relationships with validity measures were supported by significant correlations. In most cases these correlation coefficients were large (in the 0.60 to 0.90 range), demonstrating good convergent validity. All but one (TOP LIFEQ and BSI-OCD, 0.31) expected discriminant relationships were below 0.30, demonstrating excellent discriminant validity.

In many cases, there were other significant relationships between the TOP measures and validity measures of different constructs. For example, the TOP Depression measure correlated with the BASIS32-Relationship to Self and Other, and BASIS32-Daily role. As another example, the TOP Quality of Life measure correlated highly with validity scale measures of depression. One interpretation of such correlations is that many psychological constructs are not orthogonal and have been shown to inter-correlate. Another interpretation is that many psychological subscales include a portion of something like general subjective distress, which is common across different subscales.

As stated above, several TOP factors warrant further investigation. No validation subscale was used for the exact same construct measured by the TOP-Suicidal Ideation factor, although this factor demonstrated expected relationships with depression factors on the MMPI, BSI, BDI, and BASIS 32. In future investigations, this factor should be correlated with scales of suicidal ideation like the Beck Scale for Suicide Ideation (Beck, Steer, & Ranieri, 1988). Similarly, the TOP Manic factor should be correlated with other scales of mania. Although the Manic factor correlated satisfactorily with the MMPI-Hypomania scale, the items of the MMPI scale do not necessarily reflect current diagnostic classification symptoms.

Finally, the TOP Sleep and Sexual Functioning factors did not have a convergent validity measure used in this study. Both scales did show expected relationships with other related measures; however, future studies should correlate these factors with other direct measures of both of these constructs.

Study 4: Floor and Ceiling Effects

In this section, we report on the floor and ceiling effects of the TOP in a large clinical sample. Floor and ceiling effects are serious issues to consider in selecting outcome tools for clinical populations. If the tools are not able to measure the full range of pathology, their ability to accurately measure initial status and change may be severely limited. For example, Nelson, Hartman, Ojemann, & Wilcox (1995) reported that the SF-36 has significant ceiling effects in clinical samples, suggesting that the tool has limited applicability to the Medicaid population for which it was being tested. As another example, the average inpatient's Total Score at admission for the BASIS 32 is reported to be 0.79 on a scale of 0 (*no problems*) to 4 (*severe problems*) (Eisen et al., 1986). This means that the average inpatient starts near the floor of the tool and suggests that many inpatients start at the actual floor, leaving little or no room to document improvement. For the TOP to serve as a reliable and valid UCB it must demonstrate that it can measure the full range of pathology.

Method

A total of 216,642 clinical TOP administrations were analyzed for both floor and ceiling effects. Demographic information of the clinical sample is presented in Table 2. This large dataset included 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 analyze their clinical outcome data. The number of each service type is presented in Table 3. The dataset was analyzed for frequency counts of clients who scored at either the theoretical maximum or minimum score of each TOP scale. The 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. Theoretical maximum scores were calculated by scoring each measure with item scores at their highest symptom level (e.g., for the item "Indicate how much of the time during the last month you have felt down or depressed," an item score of 1 was used referring to "All of the time."). Continuing this example of depression, the DEPRS scoring results in a theoretical maximum *Z*-score of 4.63 (standard deviations from the general population mean). Similarly, the theoretical minimum score for Depression (−1.67) was calculated using the item scores representing no depressive symptoms for each item in the construct. Frequency counts were then calculated for the number of clients who actually scored at the theoretical maximum or minimum.

Results

Table 9 presents the number and percent of clients who scored at the theoretical minimum or maximum for each TOP subscale. TOP ceiling effects are virtually undetectable with only 0.1% to 4.0% of the clinical sample scoring at the theoretical maximum of TOP subscales. Only three TOP subscales had frequency counts at the maximum theoretical score greater than 1% (Quality of Life 4.0%, Sleep Functioning 2.9%, and Depression 1.1%). This result suggests that little would be gained by redesigning any subscale to have a higher maximum score.

TOP floor effects were evident on most subscales, but none of the floors are on the "pathological" side of the general population mean. In all cases the floor was below the general population mean, suggesting that each subscale is assessing the pathological range of the construct (also demonstrated by a lack of ceiling effects), but not necessarily the full "healthy" range of the construct. The most notable of floor effects occurred on Violence, Suicidality, and Sexual Functioning.

Discussion

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

Table 9
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

sample it is reasonable to conclude that the each TOP scale measures the full range of clinical severity, and represents a substantial improvement over the widely used naturalistic outcome tools reported previously.

Of particular note are the floor effects present on Suicidality and Violence. As they currently exist, both of these subscales are pathological constructs without a clear healthy side to the continuum. Having any suicidal or violent behavior is clinically defined as pathological, and is supported by the overwhelming numbers of people in the general population who do not report any problems on either of these dimensions. In other words, it is hard to report or measure less than zero suicidality. What would it mean to say that someone has an extreme score on the healthy side of violence? One generally thinks that there may be a wide range of violent thoughts, tendencies, and behaviors among people, but there is a built-in floor of little or no violent thoughts, tendencies, and behaviors, where a large percentage of the population exists. If there are any "healthy" aspects to these constructs, they are probably inoculation-type behaviors or attitudes that help insulate and protect individuals from becoming violent toward themselves or others. In the future, it would certainly be a useful goal to explore these relationships, and if they are connected to the same construct, add items to each of these measures to assist providers in not only reducing pathological behaviors, but also strengthening their resistance to these destructive actions.

Study 5: Sensitivity to Change

In this section, we report information about the TOP's 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. Ideally, evaluating sensitivity to change should include two subject samples—one that is expected to change, and another that is expected not to change based on prior knowledge or research. In addition, an external measure with proven validity and sensitivity to change should be used to verify that change has, or has not, occurred. Then the measure in question can be compared to this standard. Unfortunately, most of the constructs measured by the TOP do not have matching external measures with sensitivity to change reported in this ideal format. Therefore, less than ideal methodology must be employed.

Sensitivity to change is a critical issue for the industry to begin addressing in naturalistic settings. Many state governments (e.g. Michigan, Georgia) and private payers (e.g. Tufts) 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, Melendez, Davis, & Lunnen, 2000).

Since this is such a critical issue, if an external measure of change does not exist with proven sensitivity to change to be used as a "gold standard" of comparison, the field must not ignore this important UCB requirement. Instead, it should design studies to make the best inferences possible, allowing more informed decision-making.

Without an external "gold standard" of measurement, change documented in sensitivity to change studies cannot rule out the possibility that observed changes are the product of tool instability rather than actual change. Instead, we argue that measurement error (caused by poor reliability or validity) must be assessed prior to the study through other means (i.e., other studies of reliability and validity). First, the tool's stability should be documented (i.e., test re-test reliabilities) to ensure that change scores are not caused by errors in measurement (we have done this in Study 2). Second, the tool should demonstrate that it is effectively measuring the constructs it is supposed to be measuring (i.e., convergent and discriminant validity), which we have done in Study 3. With good test-retest reliabilities and good convergent and discriminant validity, the current study offers useful, albeit circumstantial, evidence about the TOP's sensitivity to change.

Method

Between April 1996 and June 2001, as part of routine care, 20,098 adult behavioral health clients were administered the TOP at the start of treatment and later after several therapy sessions. Age, sex, and years of education of participants are presented in Table 2 and breakdowns of service facility types are presented in Table 3. The median number of days between TOP administrations was 49 and the median treatment session at which the second TOP was administered was 7.

For each TOP subscale, within group Cohen's *d* effect sizes were calculated comparing subscale scores at first TOP administration to subscale scores at 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. 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.

Results

For each TOP subscale, Table 10 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 .16 (Mania) to .53 (Depression). The percentage of clients who made reliable improvement ranged from 10

Table 10
Sensitivity to Change

Variable	*N*	Initial Mean	Follow-up Mean	Initial *SD*	Follow-up *SD*	Cohen's *d*	% Clients Showing Reliable Improvement	% Clients Showing Reliable Worsening
DEPRS	19,660	1.34	.48	1.68	1.55	.53	54	14
VIOLN	18,765	1.25	.68	2.97	2.37	.21	31	17
SCONF	8,047	.28	−.04	1.08	1.01	.31	38	18
LIFEQ	10,039	2.19	1.44	1.83	1.81	.41	52	21
SLEEP	18,869	.68	.16	1.46	1.32	.37	47	20
SEXFN	9,407	−.12	−.31	1.12	1.04	.18	25	15
WORKF	9,600	.30	−.10	1.44	1.29	.29	39	20
PSYCS	18,320	2.02	1.14	2.85	2.42	.33	44	18
PANIC	19,701	1.36	.75	1.93	1.73	.33	41	17
MANIC	19,561	−.31	−.47	1.00	0.96	.16	10	6
SUICD	19,562	2.38	1.14	3.69	2.80	.38	42	14

(Mania) to 54 (Depression), and the percentage of clients who got reliably worse ranged from 6 (Mania) to 21 (Quality of Life). Out of 6,577 clients with scores for every subscale, 91% of clients showed reliable improvement on at least one TOP subscale and 67% of clients showed reliable worsening on at least one TOP subscale.

Discussion

Since no external measure indicating that change actually occurred was available for this study, the possibility that the TOP is unstable (rather than sensitive to change) cannot be ruled out from this study when considered in isolation. However, the strong test-retest results from Study 2 suggest that instability in the subscales is not responsible for the results from the current study. Studies 1 and 3 provide further evidence for the TOP scales' reliability and validity, suggesting that the results from the current study are not due to inaccurate measurement.

Furthermore, there is robust evidence from past research documenting the efficacy and effectiveness of psychotherapy (Feltham, 1999; Lambert & Bergin, 1994; Seligman, 1995; Shadish, 2000; Howard, Kopta, Krause, & Orlinsky, 1986; Shadish et al., 1997; Smith, Glass, & Miller, 1980). Therefore, it is reasonable to speculate that at least some of the change demonstrated in this study was real change associated with treatment rather than measurement error. However, future studies will be needed to provide definitive evidence on the issue.

This study provides evidence that the TOP may be sensitive to change. Most of the within-group Cohen's *d* effect sizes were in the small (.2) to medium (.5) range (Cohen, 1988), and may have been increased by measuring client improvement through termination. In addition, effect sizes were reported in all cases, even if the patient did not enter treatment for a problem on the dimension and already had scores at or below the general population average. This was especially true for Sexual Functioning where most patients had normal functioning at the start of treatment and had little room for, or need for improvement, on this dimension.

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.

Study 6: Criterion Validity

In this section, we report on the TOP's ability to accurately discriminate between members of the general population and behavioral health clients, and should provide further corroboration of the tentative findings discussed in Study 5. The ability of an instrument to distinguish between clients and members of the general population is important for two reasons. First, the Core Battery Conference recommended that the Universal Core Battery be able to do so as part of criterion validation. To the extent that an instrument can distinguish between clients and members of the general population, we are more likely to believe that it measures aspects of psychopathology. Second, a possible application of the TOP is to help clinicians screen potential clients to decide whether or not any treatment is needed. While the decision to treat or not should always be a matter of many factors, including clinical judgment, such decisions should be based on as much relevant information as possible, including scores on self-report tests.

Method

A total of 94 members of the general population completed the TOP. These were the same general population participants from Study 3. Demographic information of this sample is presented in Table 11 under the heading “General Population.” Age, years of education, and sex were used to create 10 unique matched samples of 94 clients each drawn from the BHL database of behavioral health clients who have taken the TOP. Binary logistic regression was applied to each set of the 94 general population participants and the matched sample from the clinical population. These analyses combined all TOP measures into a binary stepwise logistic regression to determine the most parsimonious collection of subscales accounting for independent prediction of client/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.

Results

Demographic information for the 10 client-matched samples is presented in Table 11. The extensive BHL database (more than 210,000 adult TOP administrations) allowed for very precise matching between the general population sample and the 10 sets of client samples.

In Analysis 1, the first variable entered into the model was Quality of Life, $\chi^2(1) = 40.74$, $p < .001$. Seventy percent of the clients were correctly classified as clients and 68% of general population participants were correctly classified as such, with a total classification accuracy of 69%. Psychosis was entered next, $\chi^2(1) = 10.67$, $p < .001$. With its entry, correct classification of clients increased to 73%, correct classification of the general population participants increased to 77%, and total classification accuracy increased to 75%. The results from the other four steps and the total model of analysis 1, as well as Analyses 2 through 10 are presented in Table 12.

Table 11
Demographic Information of Participants in Study 6

Analysis Number	*N*	Population	Mean Age	*SD* Age	Mean Education	*SD* Education	% Women
1–10	94	General Population	46.3	17.5	14.9	3.4	73.9
1	94	Patient	46.2	17.3	14.8	3.4	74.2
2	94	Patient	46.0	17.3	14.9	3.4	74.2
3	94	Patient	46.3	17.7	14.9	3.3	74.2
4	94	Patient	46.1	17.2	14.8	3.3	74.2
5	94	Patient	46.1	17.4	14.8	3.3	74.2
6	94	Patient	46.2	17.2	14.8	3.4	73.9
7	94	Patient	45.8	16.9	14.9	3.4	73.9
8	94	Patient	46.0	17.2	14.8	3.4	74.2
9	94	Patient	45.8	17.1	14.9	3.3	74.2
10	94	Patient	45.9	17.0	14.9	3.3	74.2

Table 12
Logistic Regression Results

Analysis No.	Step No.	Variable Entered	$\chi^2(df)$, $p <$	% Clients Classified Correctly	% General Population Participants Classified Correctly	Total % Classified Correctly	Nagelkerke R^2
1	1	LIFEQ	40.74 (1) .001	70	68	69	.28
1	2	PSYCS	10.67 (1) .001	73	77	75	.34
1	3	MANIC	14.04 (1) .001	72	80	76	.42
1	4	SUICD	9.35 (1) .01	75	83	79	.47
1	5	WORKF	7.06 (1) .01	76	86	81	.50
1	6	SEXFN	6.97 (1) .01	78	83	80	.54
1	Total Model		88.82 (6) .001	78	83	80	.54
2	1	LIFEQ	97.49 (1) .001	83	82	82	.57
2	2	MANIC	21.89 (1) .001	84	82	83	.66
2	3	DEPRS	12.91 (1) .001	86	85	86	.71
2	4	PSYCS	6.13 (1) .05	86	85	86	.73
2	5	VIOLN	5.38 (1) .05	86	85	86	.75
2	6	SEXFN	5.46 (1) .05	87	90	89	.77
2	Total Model		149.25 (6) .001	87	90	89	.77
3	1	LIFEQ	74.00 (1) .001	76	79	77	.47
3	2	MANIC	8.97 (1) .01	78	79	79	.51
3	3	PSYCS	16.07 (1) .001	82	82	82	.58
3	4	SEXFN	7.58 (1) .01	81	84	83	.62
3	5	DEPRS	4.23 (1) .05	82	86	84	.63
3	Total Model		110.84 (5) .001	82	86	84	.63
4	1	LIFEQ	73.41 (1) .001	78	79	79	.46
4	2	SEXFN	6.06 (1) .05	80	83	82	.49
4	3	PSYCS	10.39 (1) .001	78	83	80	.54
4	4	MANIC	11.44 (1) .001	83	84	83	.59
4	5	VIOLN	5.24 (1) .05	84	86	85	.61
4	6	WORKF	4.83 (1) .05	84	84	84	.63
4	7	PANIC	6.71 (1) .01	87	86	87	.66
4	Total Model		118.09 (7) .001	87	86	87	.66
5	1	LIFEQ	97.79 (1) .001	82	82	82	.58
5	2	MANIC	6.55 (1) .01	83	82	82	.61
5	3	PSYCS	14.58 (1) .001	82	84	83	.66
5	4	SEXFN	8.03 (1) .01	85	88	86	.69
5	5	PANIC	8.20 (1) .01	84	84	84	.72
5	Total Model		135.16 (5) .001	84	84	84	.72
6	1	LIFEQ	85.32 (1) .001	81	82	81	.52
6	2	WORKF	8.67 (1) .01	81	83	82	.56
6	3	PSYCS	14.24 (1) .001	82	83	83	.63
6	4	MANIC	8.73 (1) .01	83	85	84	.66
6	5	SLEEP	4.97 (1) .05	86	86	86	.68
6	Total Model		121.93 (5) .001	86	86	86	.68
7	1	LIFEQ	66.03 (1) .001	75	79	77	.42
7	2	WORKF	14.95 (1) .001	83	79	81	.50
7	3	PSYCS	13.52 (1) .001	79	80	80	.56
7	4	SEXFN	9.15 (1) .01	80	82	81	.60
7	5	VIOLN	5.78 (1) .05	80	84	82	.63
7	6	MANIC	5.37 (1) .05	83	85	84	.65
7	7	PANIC	5.73 (1) .05	85	84	84	.67
7	Total Model		120.53 (7) .001	85	84	84	.67

(*continued*)

Table 12
Continued

Analysis No.	Step No.	Variable Entered	$\chi^2(df)$, $p <$	% Clients Classified Correctly	% General Population Participants Classified Correctly	Total % Classified Correctly	Nagelkerke R^2
8	1	LIFEQ	67.35 (1) .001	76	79	78	.43
8	2	WORKF	12.67 (1) .001	74	79	76	.49
8	3	SUICD	8.74 (1) .01	76	82	79	.54
8	4	MANIC	4.45 (1) .05	77	82	79	.56
8	5	PANIC	5.18 (1) .05	82	82	82	.58
8	6	SEXFN	3.98 (1) .05	82	79	80	.60
8	Total Model		102.36 (6) .001	82	79	80	.60
9	1	LIFEQ	69.28 (1) .001	73	79	76	.45
9	2	MANIC	9.05 (1) .01	73	79	76	.49
9	3	PANIC	7.02 (1) .01	76	82	79	.53
9	4	SEXFN	6.84 (1) .01	80	83	81	.56
9	5	SUICD	5.64 (1) .05	80	83	81	.58
9	6	WORKF	6.05 (1) .05	81	83	82	.61
9	Total Model		103.88 (6) .001	81	83	82	.61
10	1	LIFEQ	67.28 (1) .001	76	79	77	.43
10	2	MANIC	9.71 (1) .01	78	78	78	.48
10	3	PSYCS	9.38 (1) .01	78	79	78	.53
10	4	SEXFN	4.99 (1) .05	80	80	80	.55
10	5	PANIC	6.90 (1) .01	82	82	82	.58
10	Total Model		98.26 (5) .001	82	82	82	.58

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. The logistic regression results for this analysis and the remaining nine analyses are presented in Table 12.

In the 10 analyses, 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 .54 to .77 with a mean of .65. In addition, the variables that were significant predictors of client/general population status were fairly consistent across the 10 analyses. In 10 of the analyses, Quality of Life and Mania were significant predictors, in 9 of the analyses Sexual Functioning was a significant predictor, in 8 of the analyses Psychosis was a significant predictor, and in 6 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 10 analyses was Quality of Life.

Discussion

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.

We were not able to find other studies with which to benchmark these results. Other criterion validity studies in the literature typically used another measure of psychopathology, the presence of a DSM diagnosis in the medical chart, or an expert rating as the criterion (Baity & Hilsenroth, 2002; Snowden, Kersten, & Roy-Byrne, 2003). Future analyses of the TOP's criterion validity should focus on larger general population samples in which all symptom and functional factors are available and a gold standard like the Structured Clinical Interview for *DSM-IV-TR* (SCID; First, Spitzer, Gibbon, & Williams, 1997) is available to accurately distinguish between groups.

General Discussion

In the present article we describe the development and initial validation of the TOP. These initial studies suggest the TOP is a promising multipurpose self-report measure. To document good psychometric properties with many different demographic and clinical populations serviced in a diverse number of treatment settings, it will be important to replicate several of the current studies that reported smaller sample sizes (especially test-retest, and convergent and discriminant validity samples). All validity and reliability studies should be replicated on diverse clinical samples to evaluate the TOP's psychometrics across the full spectrum of disorders and settings. Beyond the validity measures reported here, these future studies should include additional validity measures specifically designed for the content domains of suicidality, sexual functioning, sleep, and mania. Ideally all participants would receive all validation measures so as to assess the relative strength of correlations.

The initial results from these limited samples suggest the TOP has good test-retest reliability on all symptom and functional measures. The TOP factors correspond well with other measures of symptoms and functioning, and the TOP can distinguish between clients and members of the general population. The TOP has virtually no ceiling effects and the floor effects that do exist are not within the pathological range of the constructs. Furthermore, there is some initial evidence that the TOP subscales are sensitive to change. A definitive study of the TOP's sensitivity to change should include both a population that is expected to change and one that is not. It should also include a measure with well-documented validity and sensitivity to rule out the possibility of instability in measurement. In addition, the TOP's ability to discriminate between diagnostic groups should be tested.

The TOP-Manic scale may require additional work. Questions like "felt on top of the world" clearly are not unidimensional with respect to health, and may have very different clinical meanings for people who do, and do not, have bipolar disorder. Additional items and scoring changes may improve its internal consistency and correlation to other measures.

In summary, the self-report version of the adult TOP is a promising instrument. Its administration requires no technical expertise and typically takes only 25 minutes to complete the full battery. It surveys a broad range of symptom, functional, and case-mix variables and yields a profile of the client's condition in comparison to the general population. Good reliability and validity of the TOP and its subscales have been demonstrated with clinical and nonclinical samples.

References

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

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

Baity, M.R., & Hilsenroth, M.J. (2002). Rorschach aggressive content (AgC) variable: A study of criterion validity. Journal of Personality Assessment, 78, 275–287.

Barlow, D.H., Bach, A.K., & Tracey, S.A. (1998). The nature and development of anxiety and depression: Back to the future. In D.K. Routh & R.J. DeRubeis (Eds.), The science of clinical psychology: Accomplishments and future directions (pp. 95–120). Washington, DC: American Psychological Association).

Beck, A.T., Steer, R.A., & Garbin, M.G. (1988). Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation. Clinical Psychology Review, 8(1), 77–100.

Beck, A.T., Steer, R.A., & Ranieri, W.F. (1988). Scale for suicide ideation: Psychometric properties of a self-report version. Journal of Clinical Psychology, 44, 499–505.

Borkovec, T.D., Echemendia, R.J., Ragusea, S.A., & Ruiz, M. (2001). The Pennsylvania Practice Research Network and future possibilities for clinically meaningful and scientifically rigorous psychotherapy effectiveness research. Journal of Mental Health, 10, 241–251.

Brazier, J.E., Harper, R., Jones, N.M., O'Cathain, A., Thomas, K.J., Usherwood, T., & Westlake, L. (1992). Validating the SF-36 health survey questionnaire: new outcome measure for primary care. British Medical Journal, 305, 160–164.

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

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

Derogatis, L.R. (1975). Brief symptom inventory. Baltimore, MD: Clinical Psychometric Research.

Derogatis, L.R. (1977). SCL 90-R administration, scoring, and procedure manual. Baltimore: Clinical Psychometrics Research.

Eisen, S.V., Dill, D.L., & Grob, M.C. (1994). Reliability and validity of a brief patient-report instrument for psychiatric outcome evaluation. Hospital and Community Psychiatry, 45, 242–247.

Eisen, S.V., Grob, M.C., & Klein, A.A. (1986). BASIS: The development of a self-report measure of psychiatric inpatient evaluation. Psychiatric Hospital, 17(4), 165–171.

Eisen, S.V., Wilcox, M., Leff, H.S., Schaefer, E., & Culhane, M.A. (1999). Assessing behavioral health outcomes in outpatient programs: Reliability and validity of the BASIS-32. Journal of Behavioral Health Services Research, 26, 5–17.

Feltham, C. (Ed.). (1999). Controversies in psychotherapy and counselling. London, England: Sage Publications Ltd.

First, M.B., Spitzer, R.L., Gibbon, M., & Williams, J.B.W. (1997). Structured Clinical Interview for DSM-IV Axis I Disorders, research version, non-patient edition (SCID-I/NP). New York: New York State Psychiatric Institute, Biometrics Research.

Flynn, K. (2002). Outcome measurement in VHA mental health services. Washington, DC: Veterans Administration.

Foa, E.B., Kozak, M.J., Salkovskis P.M., Coles, M.E., & Amir N. (1998). The validation of a new Obsessive–Compulsive Disorder Scale. Psychological Assessment, 10, 206–214.

Garratt, A., Ruta, D., Abdalla, M., Buckingham, J., & Russell, I. (1993). The SF-36 Health Survey Questionnaire: An outcome measure suitable for routine use within the NHS? British Medical Journal, 306, 1440–1444.

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

Graham, J.R. (1993). MMPI-2: Assessing personality and psychopathology. New York: Oxford University Press.

Groth-Marnat, G. (1990). The handbook of psychological assessment (2nd ed.). New York: Wiley.

Hathaway, S.R., & McKinley, J.C. (1989). Manual for administration and scoring. Minneapolis, MN: University of Minnesota Press.

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.

Howard, K.I., Kopta, S.M., Krause, M.S., & Orlinsky, D.E. (1986). The dose-effect relationship in psychotherapy. American Psychologist, 41, 159–164.

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.

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.

Jette, D., & Downing, J. (1994). Health status of individuals entering a cardiac rehabilitation program as measured by the Medical Outcomes Study 36-Item Short-Form Survey (SF-36). Physical Therapy, 74(6), 521–527.

Lambert, M.J., & Bergin, A.E. (1994). The effectiveness of psychotherapy. In A.E. Bergin & S.L. Garfield (Ed.), Handbook of psychotherapy and behavior change (4th ed., pp. 143–189). New York: Wiley.

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, 247–63.

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.

Seligman, M.E.P. (1995). The effectiveness of psychotherapy: The Consumer Reports study. American Psychologist, 50, 965–974.

Shadish, W.R. (2000). The effects of psychological therapies under clinically representative conditions: A meta-analysis. Psychological Bulletin, 126, 512–529.

Shadish, W.R., Matt, G.E., Navarro, A.M., Siegle, G., Crits-Cristoph, P., Hazelrigg, M.D., et al. (1997). Evidence that therapy works in clinically representative conditions. Journal of Consulting and Clinical Psychology, 65, 355–365.

Smith, M.L., Glass, G.V., & Miller, T.I. (1980). The benefits of psychotherapy. Baltimore, MD: Johns Hopkins University Press.

Snowden, M., Kersten, S., & Roy-Byrne, P. (2003) Assessment and treatment of nursing home residents with depression or behavioral symptoms associated with dementia: A review of the literature. Journal of the American Geriatrics Society, 51, 1305–1317.

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

Trabin, T., Freeman, M.A., & Pallak, M. (1995). Inside outcomes: The national review of behavioral healthcare outcomes programs. Tiburon, CA: CentraLink Publications.

Ware, J. (1996). The MOS 36-Item Short-Form Health Survey (SF-36). In L. Sederer & B. Dickey (Eds.), Outcomes assessment in clinical practice. Baltimore, MD: Williams and Wilkins.

Ware, J.E., & Sherboume, C.D. (1992). The MOS 36-item short-form survey (SF-36): I. conceptual framework and item selection. Medical Care, 30, 473–483.

Waskow, I.E. (1975). Selection of a core battery. In I.E. Waskow & M.B. Parloff (Eds.), Psychotherapy change measures (pp. 245–269). Washington, DC: U.S. Government Printing Office.

Wells, K.B., Steward, A., Hays, R.D., Burnam, A., Rogers, W., Daniels, M., et al. (1989). The functioning and well-being of depressed patients: Results from the Medical Outcomes Study. Journal of the American Medical Association, 262, 914–919.

Copyright of Journal of Clinical Psychology is the property of John Wiley & Sons Inc. and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.