



**T e c h n o l o g y
A s s e s s m e n t
P r o g r a m**

Office of Patient Care Services

**OUTCOME MEASUREMENT
IN VHA MENTAL HEALTH SERVICES**

**An Overview and Series of
Diagnosis Specific Short Reports**

PART I. OVERVIEW:

Global measures of mental health status, psychiatric symptoms, and functioning

Author: Karen Flynn, MS, DDS
Manager, Technology Assessment Program

Contributors: Elizabeth Adams, RRT, MPH
Research Analyst

Elaine Alligood, MLS
Information Specialist

Report Released: July 2002

A SUMMARY FOR HTA REPORTS
Copyright INAHTA Secretariat 2001

VATAP is a member of the International Network of Agencies for Health Technology Assessment (INAHTA) [www.inahta.org]. INAHTA developed this checklist[®] as a quality assurance guide to foster consistency and transparency in the health technology assessment (HTA) process. VATAP will add this checklist[®] to its reports produced since 2002.

This summary form is intended as an aid for those who want to record the extent to which a HTA report meets the 17 questions presented in the checklist. It is NOT intended as a scorecard to rate the standard of HTA reports – reports may be valid and useful without meeting all of the criteria that have been listed.

OUTCOME MEASUREMENT IN VHA MENTAL HEALTH SERVICES An Overview and Series of Diagnosis Specific Short Reports PART I. OVERVIEW: Global measures of mental health status, psychiatric symptoms, and functioning (JULY 2002)			
Item	Yes	Partly	No
Preliminary			
1. Appropriate contact details for further information?	v		
2. Authors identified?	v		
3. Statement regarding conflict of interest?			v
4. Statement on whether report externally reviewed?	v		
5. Short summary in non-technical language?	v		
Why?			
6. Reference to the question that is addressed and context of the assessment?	v		
7. Scope of the assessment specified?	v		
8. Description of the health technology?	v		
How?			
9. Details on sources of information?	v		
10. Information on selection of material for assessment?	v		
11. Information on basis for interpretation of selected data?	v		
What?			
12. Results of assessment clearly presented?	v		
13. Interpretation of the assessment results included?	v		
What Then?			
14. Findings of the assessment discussed?	v		
15. Medico-legal implications considered?			v

OUTCOME MEASUREMENT – MENTAL HEALTH OVERVIEW: FINAL REPORT

16. Conclusions from assessment clearly stated?	v		
17. Suggestions for further actions?	v		

ACKNOWLEDGMENTS

The VATAP gratefully acknowledges the contributions of the following reviewers. Participation as a reviewer does not imply endorsement.

Valerie A. Lawrence, MD
Physician Advisor, VA Technology Assessment Program
Audie L. Murphy VA Medical Center
Associate Professor, Dept of Medicine
University of Texas Health Science Center
San Antonio, Texas

Thomas Holohan, MD
Chief Officer
Office of Patient Care Services
Veterans Health Administration
Washington, DC

Laurent Lehmann, MD
Chief Consultant, Mental Health Strategic Healthcare Group
Office of Patient Care Services
Veterans Health Administration
Washington, DC

TABLE OF CONTENTS

EXECUTIVE SUMMARY i

ABBREVIATIONS COMMONLY USED THIS REPORT iii

I. INTRODUCTION..... 1

 A. *Purpose* 1

 B. *Background*..... 1

II. ORGANIZATION OF THIS REPORT 2

III. BURDEN OF DISEASE..... 3

 A. *Definitions of “serious mental illness”*..... 3

 B. *Impact of serious mental illness on patients’ lives*..... 3

 C. *VHA mental health care populations*..... 3

IV. VHA MENTAL HEALTH SERVICES..... 4

V. METHODS FOR THE SYSTEMATIC REVIEW 4

 A. *Selection criteria (for standardized measures of mental health care outcomes to be used within VHA)*..... 5

VI. RESULTS 6

 A. *Standardized measurement instruments that met selection criteria*..... 6

 B. *Parallel reviews conducted for other national health care systems: results, recommendations, and implementation suggestions*..... 9

VII. SUMMARY AND DISCUSSION..... 9

Table 1. Measures Evaluation Matrix: Abstracted Details For Global Mental Health Status, Psychiatric Symptoms, Functioning, Disabilities Measures..... 11

Table 2. Measures Evaluation Matrix Summary: Global Mental Health Status, Psychiatric Symptoms, Functioning, Disabilities Measures (Specific To Mental Health, Not General Health Or Emotional Health Secondary To A General Medical Condition) 27

VIII. REFERENCES 21

EXECUTIVE SUMMARY

‡ **Background and VHA policy requirements:** The US Veterans Health Administration (VHA) is accountable for the quality of mental health services provided to its patients. In this report the VHA Technology Assessment Program (VATAP) reviews available standardized mental health care outcome measurement instruments to assist in identifying those most suitable for routine outcome data collection and reporting by VHA. VHA’s policy needs dictate that instruments be reliable and valid, applicable to serious mental illness (schizophrenia, major depression, substance abuse, and post traumatic stress disorder), and sensitive to change after treatment. Instruments should be feasible for routine clinical use, i.e., imposing minimal respondent or interviewer burden, and brief. Finally, cost for VHA to obtain and use should be minimal and electronic data collection and analysis should be available.

‡ **Value added by VATAP to existing work:** VATAP searches of electronic literature databases identified existing compendia of standardized mental health care outcomes measures and reviews produced for other national health care systems seeking to implement routine outcome data collection in mental health services. Notable among the former is the American Psychiatric Association’s *Handbook of Psychiatric Measures*, and among the latter reviews from Australia and the UK. The large number of existing compendia and reviews allowed VATAP to **add additional levels of synthesis and VHA relevance to existing work:**

‡ **Review of the psychometric literature focused on VHA mental health care outcomes measurement needs:** VATAP’s review of the literature on developing and testing standardized health status measurement instruments is reported in the Appendix. The review led VATAP to generate a list of selection criteria for measures for use in VHA mental health services. These criteria are discussed fully in the Methods section and include:

- ?? original purpose of instrument congruent with VHA intended use;
- ?? multi-dimensional (measuring both symptoms and patient functioning);
- ?? acceptable reliability and validity;
- ?? sensitive to change with treatment;
- ?? feasible for routine use;
- ?? electronic version available;
- ?? yield data that are readily interpretable by a wide audience;
- ?? free or obtainable by VHA at low cost.

‡ **Global AND diagnosis-specific instruments are addressed:** This “Overview” reports on VATAP’s application of those criteria to global measures of mental health status, psychiatric symptoms, and functioning. It is accompanied by a series of diagnosis-specific reports.

‡ **VHA’s choice of global versus diagnosis-specific instruments:** The literature offers expert opinion indicating that a combination of global and diagnosis-

specific measures may be the most flexible approach: a global scale used in conjunction with clinically useful ones uniquely suited to particular target groups can serve several functions, such as reporting to stakeholders, and documenting treatment effectiveness for clinical planning.

‡ **Selection of measures for VHA should recognize context and organizational needs:** Final selection of an instrument (or battery of instruments) for VHA use will require agreement among policy makers, administrators, clinicians, and stakeholders regarding preferred constructs and reporting formats for VHA mental health services quality.

‡ **Further field-testing of an instrument already used within VHA may be warranted:** In the interim, VHA should consider field-testing selected measures on this report's short list of five. The list includes the Global Assessment of Functioning (GAF), Health of the Nation Outcomes Scales (HoNOS); Brief Psychiatric Rating Scale (BPRS) Behavior and Symptom Identification scale-12 item (BASIS-12); Threshold Assessment Grid (TAG); and Compass Out-Patient (Compass-OP). The literature confirms that GAF is one of the most commonly used mental health care outcomes measurement instruments. Familiarity with, and existing investment in, GAF would argue for its serious consideration.

‡ **Changing opinions on GAF within VHA:** While clinical experience with GAF within VHA has not been uniformly positive, recent research findings and training activities have increased the probability of its acceptance and use as the basis for large-scale data collection efforts.

‡ **VATAP will continue to monitor the mental health care outcome measurement literature on behalf of VHA:** Instruments additional to those available for this report are under development.

ABBREVIATIONS COMMONLY USED THIS REPORT

AHRQ	Agency for Healthcare Research and Quality
APA	American Psychiatric Association
BASIS-32	Behavior and Symptom Identification Scale
BPRS	Brief Psychiatric Rating Scale
BSI	Brief Symptom Inventory
CAN	Camberwell Assessment of Need
CGI	Clinical Global Impressions scale
COMPASS-OP	Compass Out-Patient
GAF	Global Assessment of Function
HoNOS	Health of the Nation Outcomes Scales
JCAHO	Joint Commission on Accreditation of Healthcare Organizations
MOT	Medical Outcomes Trust
TAG	Threshold Assessment Grid

**OUTCOME MEASUREMENT IN VHA MENTAL HEALTH SERVICES
PART I. OVERVIEW:
GLOBAL MEASURES OF MENTAL HEALTH STATUS, PSYCHIATRIC SYMPTOMS,
AND FUNCTIONINGS**

I. INTRODUCTION

A. Purpose

This VA Technology Assessment Program (VATAP) report was written in response to a request by the VHA to assist in identifying which standardized mental health care outcome measurement instruments are suitable for routine outcome data collection and reporting by the VHA. VHA's policy needs dictate that instruments be reliable and valid, applicable to serious mental illness (schizophrenia, major depression, substance abuse, and post traumatic stress disorder), and sensitive to change after treatment.

B. Background

The problems associated with measuring and reporting on mental health care quality are discussed at length in the literature (Lohr, 1992).

Before a discussion of outcome measurement instruments can occur, we must agree on a definition of mental health care outcome. "Outcome" for mental health care purposes can be defined in several ways. Andrews (1994), reporting on a review of mental health care outcome measures for routine use in the Australian national health care system, defined outcome simply: "the effect on a patient's health status attributable to an intervention by a health professional or health service". Lohr (1988) cites outcome as one of Donabedian's classic triad of components of health care quality and explores some of the connections among the components still in need of research definition:

"Patient outcomes form one part of the classic triad used to define quality of care: structure, process, and outcome. Outcomes are the end results of medical care: what happened to the patient in terms of palliation, control of illness, cure, or rehabilitation. The concept of outcome directs attention specifically to the patient's well-being; it emphasizes individuals over groups, and the interests of unique patients over those of society..."

... Outcome measurement – a central concept of quality of care – has both conceptual appeal and limitations as a practical assessment tool. The degree to which outcomes can be directly related to processes of care continues to be especially problematic... To strengthen our understanding of both (process and outcomes) measures in ascertaining quality of care, I suggest that work in four areas is needed: more definitive evidence of process and outcome linkages; stronger relationships between technology assessment and quality assessment; improved reliability and validity of

outcome measures as screening tools; and continued development of health status measures.” (Lohr, 1988)

Lohr emphasizes the individual. This is in agreement with VHA’s patient focus, i.e., demonstrating patient satisfaction and other outcomes of care from the patient perspective. Lohr’s definition also accommodates the system perspective and goal of maximizing efficient use of limited resources for effective and clinically relevant interventions. For the purposes of this report, VATAP will rely on the definition of outcome proposed by Lohr (1988).

The outcome in which VHA is most interested is change with treatment. VHA’s policy needs indicate that surrogate, process-oriented measures of outcome, such as changes in frequency of outpatient appointments or time-to-follow-up are inadequate. Hence, these measures will not be included in this report.

Many standardized mental health care outcome measurement instruments are available. The most commonly used instruments are the Global Assessment of Function (GAF), and the Brief Psychiatric Rating Scale (BPRS).

Jenkins (1990) reported dissatisfaction with many of the outcome indicators currently in use such as mortality, morbidity, disability days, bed days, restricted activity days, hospital admission figures and subjective health indicators. Jenkins (1990) concluded that direct measures of health and social functioning *“have by far the strongest conceptual basis of relevant indicators of health outcome”*.

VHA shares responsibility for data collection and reporting on quality of care, including outcomes of mental health care, with other large health care systems internationally (Slade, 1999a; Brugha, 1996; Stedman, 1997; Barry, 1997; Clifford, 1999) and nationally (Slutsky, 2001; Pallak, 1994). For this reason, the system of outcome measurement must be as accurate as possible.

Psychiatry is turning increasingly to reliance on research evidence as a basis for clinical and policy decision (Powell and Geddes, 1997). VHA as a system also has embraced evidence-based decision making. A core component of the evidence paradigm, as well as that of quality, is outcome measurement. This further emphasizes the need for mental health instruments to be reliable, valid and applicable to serious mental illness such as schizophrenia, major depression, substance abuse, and post traumatic stress disorder.

II. ORGANIZATION OF THIS REPORT

This “Overview” discusses standardized global measures of mental health status, psychiatric symptoms, and functioning. It is accompanied by a series of diagnosis-specific “short reports” (a standard VATAP format), each focused on existing instruments for assessing outcomes of treatment for one of the serious mental illnesses of concern to VHA. Diagnosis-specific instruments are evaluated by the same criteria applied to global instruments.

The entire assembly of reports is accompanied by an Appendix. It comprises a glossary of psychometric terms combined with a discussion of health status measurement issues. VATAP used the background material in the Appendix to generate provisional selection criteria for measures to be used within VHA as the basis of routine outcome data collection for its mental health services.

III. BURDEN OF DISEASE

A. Definitions of “serious mental illness”

The fourth edition of the psychiatric Diagnostic and Statistical Manual (DSM-IV) defines “mental disorders” as a broad range of conditions characterized by abnormal behavioral and psychological signs and symptoms that result in dysfunction. Details on the diagnoses specified for this review are provided in the report sub-sections under each diagnostic heading.

Severe mental illness has been operationally defined by several criteria sets (Ruggeri, 2000; Huxley, 2000), some of which set threshold scores on standardized instruments as one among the criteria. For the purpose of this report, diagnoses considered “serious” were specified by the DUSH in her written request.

B. Impact of serious mental illness on patients’ lives

Serious mental health disorders interact in multiple complex ways with other aspects of health, life, and function. Andrews (1994) provided the following description:

*“People with mental disorders have symptoms and behaviors that can impair their ability to work and love, and that can impair access to physical health care, income maintenance, education, housing, transport, legal advice, and leisure opportunities. While each citizen, mentally ill or not, has the right to have these commodities, citizens with mental disorders will often find access reduced, specifically because the symptoms and behaviors associated with the mental disorder impair their ability to compete for access. Any measure of outcome...must encompass changes in symptoms, **and** changes in disabilities in the above areas.”*

Thus, optimal measures for outcome in mental health care can be expected to be correspondingly multi-dimensional, addressing domains of symptoms, perceived needs, functional disabilities, and quality of life. This is in agreement with the VHA’s preference for multidimensional measures (Holohan and Lehmann, 2001).

C. VHA mental health care populations

Jansen (2001) reports that in FY 2000, VHA treated 678,932 patients in specialized mental health care programs, and a further 153,474 in non-specialized mental health care, making a total of 832,406 patients treated within VHA for mental disorders, or 23%

of all patients treated in VHA in 2000. The proportion of veterans treated for mental health disorders annually increased by 37% between 1994 and 2000.

IV. VHA MENTAL HEALTH SERVICES

(VHA Program Guide 1103.3, 1999)

VHA's mission, as stated by the Under Secretary for Health in 1997, is: *"to serve the needs of America's veterans by providing primary care, specialized care, and related medical and social support services..."*

This mission is reflected in VHA mental health services through:

1. An integrated continuum of mental health services;
2. A shift from an inpatient focus to residential treatment and community-based services;
3. Integration of specialized mental health knowledge into primary care;
4. An orientation to quality that stresses maximizing each patient's functional independence, hence providing care in the most appropriate location according to medical condition and functional status;
5. Evaluation of mental health services that includes outcomes monitoring and benchmarking with comparable non-VHA health care systems.

V. METHODS FOR THE SYSTEMATIC REVIEW

A computerized search of the psychological and biomedical databases, MEDLINE[®], HEALTHStar[®], PSYCInfo[®], Current Contents[®], EMBASE[®], The Cochrane Library[®], and local monograph collections (McLean Hospital, Belmont, MA, and the Countway Library of Medicine at Harvard University) was conducted for the time period between 1976 and 2001.

Bibliographic search strategy terms included many exploded MeSH[®] subject headings for the serious mental illnesses: schizophrenia, psychosis, depression, PTSD, substance abuse, and the text word phrase 'serious mental illness.' These terms and their synonyms combined with terms and free text words describing treatment outcome, outcome measures, outcome assessment, treatment efficacy, and outcome evaluation yielded substantial results. These results were combined with additional terms describing study designs, types, randomization, systematic reviews, and age groups (adult, middle age, aged).

The strategies outlined above and review of end references produced over 1400 references. The enormous size of this literature resulted in the use of compendia rather than original research reports, with reference to the latter where needed or helpful. This

resulted in the use of approximately 300 papers. Of these, the *Handbook of Psychiatric Measures* [American Psychiatric Association (APA), 2000] and existing reviews similar in purpose to that needed by VHA were used (e.g., Andrews, 1994; Slade, 2001).

Thus, the present report is tailored specifically to VHA's needs and mission. This review can be considered a secondary review of other resources, supplemented from the primary research literature as needed, and framed in the context of VHA mental health services.

A. Selection criteria (for standardized measures of mental health care outcomes to be used within VHA)

Methods for developing and evaluating health status measurement instruments indicated significant agreement on prerequisites for the selection of measures for routine use within health care systems (Lohr, 1996; Sackett, 1977; Switzer, 1999; Zarin, 2000; Smith, 2000; all contributing to the Appendix). This is in agreement with a UK psychiatric services researcher currently engaged in a similar task to the VATAP review (Slade, 2001), and reports identifying standardized measures for routine use in tracking outcomes of mental health care conducted for the Australian national health care system (Andrews, 1994; Stedman, 1997).

From this agreement, and in the context of VHA mental health services as described in VHA Program Guide 1103.3 (1999), VATAP generated a list of instrument selection criteria relevant to VHA:

1. Original purpose congruent with intended VHA use for quality of care tracking and reporting, documenting effective treatment of veteran mental health patients treated in a variety of settings;
2. Multidimensional (e.g., covering both symptoms and functioning);
3. Acceptable psychometrics;
4. Sensitivity to change;
5. Feasible for routine use (i.e., brief, imposing minimal administrative, clinical, and respondent burdens);
6. Electronic data entry and/or analysis available;
7. Readily interpretable by non-professionals (i.e., summary score of generally understood construct not requiring extensive background information for interpretation and immediate understanding);
8. Low cost (i.e., instrument in the public domain or available for distribution and use at minimal cost to VHA).

Application of these criteria to global standardized instruments listed in the APA's *Handbook of Psychiatric Measures* (2000) is detailed in Table 1 and summarized in Table 2. The measures included in the tables are those listed in the *Handbook's* chapters: "Choosing, Using, and Interpreting Measures for Health Care Systems"; "General Psychiatric Symptoms Measures"; "Mental Health Status, Disabilities, and Functioning". When VATAP identified new or additional instruments not listed in the

APA *Handbook*, they were included in Tables 1 and 2, with citations from their original research literature.

A high quality review (Barry, 1997) of the applications of quality of life (QOL) measures to patients with serious mental illness found such measures inappropriate for use as routine outcome indicators. This is in agreement with Andrews (1994) and Greenley (1997). As a result, VATAP eliminated QOL measures from consideration as potential routine outcome indicators for VHA mental health services. Instead, the focus of this overview is global measures of psychiatric symptoms, mental health status, and functioning.

VI. RESULTS

After abstracting information on standardized global symptoms and functioning instruments from the *APA Handbook of Psychiatric Measures* (2000) and other sources in Table 1, the same information was further summarized according to criteria for selection of measures for VHA use in Table 2.

Table 2's shaded rows indicate that five standardized instruments appear to meet most of VHA's needs. Strengths and weaknesses of each instrument on the "short list" generated by this review, which may not be immediately apparent from Table 2, are further detailed below.

Two additional measures, the Camberwell Assessment of Need (CAN) and Threshold Assessment Grid (TAG) meet seven of the eight criteria, in both cases lacking only electronic versions for data entry and analyses. As they are otherwise promising for brevity and coverage of relevant domains, copies are included in the Appendix. TAG is further included in the discussion below, due both to its promising characteristics and presence on the short list of Slade's (2001) analogous review.

A. Standardized measurement instruments that met selection criteria

Global Assessment of Functioning (GAF)

Reports indicate that this measure met the criteria for use in the VHA. However, informal discussions with VHA clinicians suggest it did not accurately reflect severity of illness in diagnoses prevalent among VHA patients, such as post-traumatic stress disorder (PTSD).

Formal training by means of video and satellite conferencing has increased appreciation of and enthusiasm for GAF within VHA. In this context and with the strengths apparent from Tables 1 and 2, GAF deserves continued use, at least on an interim basis, within VHA mental health services. Recent research using GAF within VHA indicates that the GAF is sensitive to change and may be useful in risk-adjusting mental health outcomes across VHA (Rosenheck and Van Stone, 2001).

Coffey (1996) found that while GAF was acceptable for routine clinical use, it was also unsuitable for making individual treatment decisions. Accordingly, Coffey (1996)

suggests that the role of GAF might be that of a first step before using more sophisticated scales:

“The GAF may be adequate at the macro level and could be used for comparative studies with other services or models of services. It may also be useful as a measure of morbidity within a service or as one of a series of measures in the costing of services.”

Coffey’s statement may be interpreted as supporting use of GAF as a first-step test. In that case, assuming maximal sensitivity, negative (low scores) would be true negatives, and perhaps certain threshold scores could be used to discharge patients from particular levels of care to the next less restrictive level. The more sophisticated scales used could be the other instruments on this review’s short list, or the diagnosis-specific instruments enumerated in the corresponding sections of this review.

Piersma and Boes (1997) report that GAF is probably the most commonly used clinician rating scale to assess level of dysfunction, making it the *de facto* choice for an outcome measure in many organizations. In this context, VHA may find itself in a good position for “benchmarking” with other health care systems. The same authors report sensitivity to change data that may enhance VHA’s comfort with GAF.

Health of the Nation Outcome Scales (HoNOS)

HoNOS was developed to measure mental health care outcomes in response to government-set goals for a publicly financed national health care system (the UK National Health Service). The goals of HoNOS developers were: brevity and feasibility for routine use; coverage of common clinical problems and social functioning; sensitivity to change; reliability; and correlation with established scales (Sharma, 1999). This use is similar to that of VHA.

Sharma (1999) evaluated HoNOS in routine clinical practice. This study found a reduction in HoNOS scores after 6 months of treatment in 204 patients in general psychiatric practice (Liverpool, UK). The change in HoNOS scores was consistent with that measured concurrently with the Clinical Global Impressions Scale (see Table 1). Sharma concluded that HoNOS could be administered during routine patient care.

However, widespread implementation of the scales would be premature. Patients with disorders other than psychotic and affective showed little score change with treatment; and the data were of limited value in care planning in daily practice. In addition, a field trial of HoNOS in the Australian state of Victoria (Trauer, 1999) found unreliability in certain scale items (accommodation and occupation), an inconsistent sensitivity to change, and also an inconsistent relationship between clinically perceived severity of illness and HoNOS score.

Vagaries of regional English could further discourage VHA’s use of the scales. However, close inspection of the scales by multi-lingual (for varieties of English) VATAP staff failed to reveal overt examples of British usage likely to confound speakers of standard 20th century American English.

HoNOS is free for use in the UK, but would require permission from the Royal College of Psychiatrists for use in VHA, adding again to questions regarding potential VHA acceptability. Finally, both start-up and periodic follow-up training is recommended, a time and resource commitment that could further complicate acceptability of HoNOS for VHA use.

Behavior and Symptom Identification Scale (BASIS-32)

This instrument was developed in the US for use with inpatients. It has since been tested for reliability with outpatients (Eisen, 1999). These authors, the instrument's developers, concluded that further data from a wide range of facilities and patient samples would further validate and refine it. BASIS-32 is brief, can be completed by non-professionals, and could likely be used free with permission within VHA.

However, concerns with the instrument's psychometrics have been reported. Blais (1999) cites studies demonstrating that three of the five BASIS-32 sub-scales are highly interrelated, suggesting that they may be measuring a single dimension rather than different aspects of functioning.

Brief Psychiatric Rating Scale (BPRS)

This scale has been widely used in efficacy studies and to validate other scales (Hafkenschild, 2000). On these grounds, it should be familiar to VHA clinicians. It is in the public domain, and the time needed to complete it is comparable to that for other instruments on this review's short list. However, should VHA opt for a single global outcome instrument in preference to a battery including diagnosis-specific instruments, BPRS is sensitive to change only for patients with high symptom levels. Therefore, it may be increasingly less useful as patients improve, but still warrant quality tracking and reporting.

Threshold Assessment Grid (TAG)

This instrument is very brief and in the public domain, but its development and reliability testing remain incomplete (Slade, 2001). Currently available reports do not address availability of electronic data entry or analysis. Otherwise it is well suited to routine use within VHA.

Compass out-patient (COMPASS-OP)

This instrument is relatively brief, but is also copyrighted, and precise costs to use it are not noted in the *APA Handbook*. Finally, it was developed for outpatient use and may be less sensitive to change in more severely ill in-patients.

B. Parallel reviews conducted for other national health care systems: results, recommendations, and implementation suggestions

Andrews (1994)

This review of mental health care outcomes measures for the Australian health care system recommended field-testing four measures potentially suitable for routine use:

- ?? Behavior and Symptom Identification Scale (BASIS-32)
- ?? Health of the Nation Outcomes Scales (HoNOS)
- ?? Medical Outcomes Study Short Form (SF-36)
- ?? Mental Health Inventory (MHI)

Stedman (1997) reports on the field testing of the scales recommended by Andrews (1994), and concludes that *“scales which assess consumer outcomes are feasible for use in the settings and groups included in the project. Simple global assessments of change, and visual analog scales also included in the project, may have a place for certain purposes. Strategies for implementation of any routine outcome measures should include training and education, consultation, a multi-disciplinary approach, and research (to increase understanding of how to measure change and how to interpret the findings in different situations)”*.

Slade (2001)

This author’s confidential (unpublished) review used criteria analogous to this VATAP review. Slade’s short list of instruments for potential routine use included:

- ?? Threshold Assessment Grid (TAG)
- ?? Manchester Short Assessment of Quality of Life (MANSA)
- ?? Helping Alliance Scale (HAS)
- ?? Camberwell Assessment of Need Short Assessment Schedule (CANSAS)

Slade notes that other instruments suitable for routine use are under development and testing and include the Dartmouth COOP Functional Assessment Charts, Helping Alliance Scale, Threshold Assessment Grid, and Functional Assessment of Care Environments core assessment. VATAP concurs that instruments designed to support routine use warrant monitoring of subsequent research results, specifically with respect to their psychometric properties and acceptability in pilot tests of routine use. VATAP will continue to monitor other newly available resources relevant to VHA’s needs.

VII. SUMMARY AND DISCUSSION

VATAP’s review process has generated a list of five standardized mental health care outcomes instruments that meet most criteria for routine use within VHA. These instruments are: the Global Assessment of Functioning (GAF), Health of the Nation Outcomes Scales (HoNOS); Brief Psychiatric Rating Scale (BPRS) Behavior and Symptom Identification scale-12 item (BASIS-12); Threshold Assessment Grid (TAG); and Compass Out-Patient (Compass-OP). However, VHA mental health clinicians’ mixed experiences with at least one instrument (GAF) on the list would argue for VHA

field trials and pilot testing of several of the short-listed instruments for clinician acceptability prior to final instrument selection and implementation of large-scale data collection efforts.

Slutsky (2001) reported that implementation of an Agency for Healthcare Research and Quality (AHRQ) Clearinghouse for Quality Indicators analogous to the existing Guidelines Clearinghouse is imminent; the Quality Clearinghouse also will be searchable via the Internet. AHRQ has funded a proposal to catalog mental health care indicators for the Clearinghouse (M. Erlichman: personal communication 2001). Also warranted would be continuing VATAP communication with colleagues engaged in similar review efforts. The Clearinghouse and ongoing communication will allow tracking of results for instruments currently under development or field testing, in the expectation that some newer instruments will be more ideally suited to VHA's purposes in routinely measuring outcomes for its mental health services.

TABLE 1. Measures Evaluation Matrix: Abstracted Details For Global Mental Health Status, Psychiatric Symptoms, Functioning, Disabilities Measures

Notes: Unless otherwise noted, the information in this table was obtained from The American Psychiatric Association’s *Handbook of Psychiatric Measures* (2000). Instruments still under development were omitted from this and Table 2, as available information is likely to be incomplete. Shaded rows indicate that five standardized instruments appear to meet most of VHA’s needs.

Criteria for VHA use of standardized measures of mental health care outcome:

1. Original purpose congruent with intended VA use for quality of care tracking and reporting, documenting effective treatment of veteran patients
2. Multidimensional
3. Acceptable psychometrics: reliability; validity
4. Sensitive to change
5. Feasible for routine use: time; acceptability; who administers
6. Electronic data entry, analysis
7. Readily interpretable by non-professionals
8. Free or obtainable to VA at minimal cost

Name	Purpose	Multi-Dimensional?	Reliability	Validity	Change?	Time	Acceptability	Who?	Electronic?	Interpretation	Cost
------	---------	--------------------	-------------	----------	---------	------	---------------	------	-------------	----------------	------

OUTCOME MEASUREMENT – MENTAL HEALTH OVERVIEW: FINAL REPORT

Name	Purpose	Multi-Dimensional?	Reliability	Validity	Change?	Time	Acceptability	Who?	Electronic?	Interpretation	Cost
GAF	Yes, can be made to fit	Yes	0.61-0.91 (fair to excellent)	Good concurrent validity, but depends on info available to guide rating May confound symptoms and functioning	More sensitive than other measures	Once info obtained, 1-2 minutes	Already in use	Clinician, with structured interview, guidelines, users' guide	No	Summary score of functioning	Public domain - free

OUTCOME MEASUREMENT – MENTAL HEALTH OVERVIEW: FINAL REPORT

Name	Purpose	Multi-Dimensional?	Reliability	Validity	Change?	Time	Acceptability	Who?	Electronic?	Interpretation	Cost
HoNOS	Yes, for routine use in large publicly funded system to record and measure problems and to track change over time	Yes	Very good	Good	Sensitive to Improvement and deterioration	15-30 minutes first time, less on subsequent administrations	4 hour training session recommended, with subsequent 4 hour training session recommended, with subsequent supervision to assure reliability good acceptability in field tests	Clinician, after routine clinical interview:	Yes	Summary score	Materials and manuals, \$20 Permission to copy must be obtained from Royal College of Psychiatrists Research Unit

OUTCOME MEASUREMENT – MENTAL HEALTH OVERVIEW: FINAL REPORT

Name	Purpose	Multi-Dimensional?	Reliability	Validity	Change?	Time	Acceptability	Who?	Electronic?	Interpretation	Cost
BASIS-32	<p>?Brief, comprehensive measure of mental health treatment outcome from patient's perspective</p> <p>?Broad measure of general functioning that can evaluate change over course of treatment</p> <p>?Behavior, functioning, and psychiatric symptoms</p>	Yes	<p>?Cronbach's alpha for subscales =.65-.81</p> <p>?Full scale internal consistency = .89</p> <p>?Test-retest reliability coefficient = .65-.81</p> <p>?Some questions cited in literature are high interrelation of subscales</p>	<p>?Multiple validation studies support use in psychiatric in-patients and outpatients</p> <p>?MOT listed</p> <p>?JCAHO – approved performance measurement</p>	yes	<p>?Self-report: 5-20 minutes</p> <p>?Structured interview: 15-20 minutes</p>	Good: simple, brief, can be administered by non-professionals	Self-report, structured interview by professional or non-professional	Yes, several versions available from commercial vendors	Sub-scale and overall scores	Copyrighted, but holder has given mental health professionals permission to reproduce manual version and track outcomes in own patients

OUTCOME MEASUREMENT – MENTAL HEALTH OVERVIEW: FINAL REPORT

Name	Purpose	Multi-Dimensional?	Reliability	Validity	Change?	Time	Acceptability	Who?	Electronic?	Interpretation	Cost
BPRS	Change in severity of symptoms of psychopathology (symptom change in patients with psychotic illness) ?Effectiveness of treatment ?To classify patients into subgroups ?To summarize patient characteristics that may predict treatment response		Varies with training and experience of clinician rater: Good joint reliability requires considerable time and effort (results of joint rating sessions discussed to improve reliability)	Inpatients and outpatients, but less useful in patients with low levels of psychopathology	Only in patients with high levels of psychopathology, less so in patients with lower levels	20-30 minutes, depending on familiarity with patient and on patient cooperation	Often used to assess effectiveness of interventions. Clinical use less well-documented, and does not cover all areas of potential clinical interest.	Not reported by APA	Experienced clinician, using information from interview and patient observation	Some subscales may have descriptive utility; groupings of positive and negative symptom items may have good reliability	Public domain Semi-structured interview guides are available

OUTCOME MEASUREMENT – MENTAL HEALTH OVERVIEW: FINAL REPORT

Name	Purpose	Multi-Dimensional?	Reliability	Validity	Change?	Time	Acceptability	Who?	Electronic?	Interpretation	Cost
COMPASS-OP	To assess: ?General mental health status ?Satisfaction with care ?Patient characteristics ?Outcomes of care ?Some elements of the process of care ?For patients with any mental health disorder in outpatient care ?Health care delivery system performance		?Internal consistency very good ?Test-retest high	Good correlation with GAF, SF-36, MHI, SCL-90-R, other depression scales for relevant sub-scales	Yes	?25 minutes for patients, 5 minutes for clinicians at initial visit ?15 minutes for patients, 3 minutes for clinicians at each follow-up	Probably good, although developed for outpatients and may be less appropriate in populations with more severe conditions. Training recommended	Clinician and patient	Yes	Mental health index (MHI) provides overall aggregate score by summing 3 sub-scales outcomes calculated by difference between observed and expected, adjusted for norms	Copyrighted, fees for use to be negotiated

OUTCOME MEASUREMENT – MENTAL HEALTH OVERVIEW: FINAL REPORT

Name	Purpose	Multi-Dimensional?	Reliability	Validity	Change?	Time	Acceptability	Who?	Electronic?	Interpretation	Cost
BSI	Symptom inventory for respondents in community, medical (primary care) and psychiatric settings, derived from SCL - 90		Moderate to good	Demonstrated to be broadly sensitive to manifestation of psychological distress across a wide range of contexts	Not documented by APA	2-5 minutes of instruction, 8-10 minutes to self-complete	Effective screening tool for identification of psychiatric disorders in primary care	Self-report (Clinician versions also available)	Yes	Same as SCL-90-R	Copyrighted, APA does not provide cost

OUTCOME MEASUREMENT – MENTAL HEALTH OVERVIEW: FINAL REPORT

Name	Purpose	Multi-Dimensional?	Reliability	Validity	Change?	Time	Acceptability	Who?	Electronic?	Interpretation	Cost
CAN (Slade, 1999b)	To measure the needs of people with severe and enduring mental illness. Clinical (care planning) and research/ (outcome, service evaluation) versions differ in only one section	Yes	Inter-rater and test-retest very high	Face and content validity demonstrated. Aggregate score of 7 needs correlates with GAF score	Not specifically noted	?25 minutes ?short form: 3-5 minutes		Variety of professionals, no special training.	?Developed but found not feasible to use ?SPSS templates can be downloaded from Internet		TAP purchased book, from which CAN allowed to be duplicated "freely"

OUTCOME MEASUREMENT – MENTAL HEALTH OVERVIEW: FINAL REPORT

Name	Purpose	Multi-Dimensional?	Reliability	Validity	Change?	Time	Acceptability	Who?	Electronic?	Interpretation	Cost
CGI	Yes: 3 indices: ?Severity ?Improvement ?Efficacy	Yes	Mixed research results: ?Test retest reliability low ?Internal consistency high	Good concurrent validity	Good sensitivity to change over time	1-2 minutes after clinical interview	Widely used outcome scale in psycho-Pharmacology trials in spite of mixed results on reliability	Clinician	No		Public domain
GHQ	No, assessment of psychiatric distress related to general medical illness, screening for formal psychiatric interview										

OUTCOME MEASUREMENT – MENTAL HEALTH OVERVIEW: FINAL REPORT

Name	Purpose	Multi-Dimensional?	Reliability	Validity	Change?	Time	Acceptability	Who?	Electronic?	Interpretation	Cost
MCAS	No, designed for people with chronic mental illness living in community, so may not be relevant to other settings		Test-retest good	Good	Not reported	Few minutes	Describes case mix, facilitates treatment planning, tracks patient progress, developed for restricted range of patients	Case manager	Nor reported		Public domain
MHI (Both 38 and 18 item versions)	No, intended to assess mental health in psychiatrically healthy samples distress related to general medical illness, not appropriate for psychiatric populations		??83-.91 for scales (lower order factors) ??92-.96 (higher order factors) ??Test-retest, .56-.64 ??Internal consistency good	Better than “standard” instruments in comparison studies for detecting affective or anxiety disorders							

OUTCOME MEASUREMENT – MENTAL HEALTH OVERVIEW: FINAL REPORT

Name	Purpose	Multi-Dimensional?	Reliability	Validity	Change?	Time	Acceptability	Who?	Electronic?	Interpretation	Cost
MHQ	Rapid quantification of symptoms and traits of psychoneurotic illness and personality disorder Use in clinical research, psychiatry practice, general medicine, industry, education	Yes	Some items may tap dimensions not uniquely associated with their sub-scales	Mixed support for sub-scale content (particularly the hysteria sub-scale)	Some studies suggest usefulness as a measure of change that is sensitive to treatment effects, but more widely used in research and epidemiology studies	5-10 minutes for self completion	Designed for brevity and simplicity, but APA reports that sub-scales don't discriminate well between corresponding diagnoses or clinicians' ratings	Self-report by patient	No	General measure of subjective psychological distress across a wide range of demographics and diagnostic groups	Copyrighted, cost not noted by APA

OUTCOME MEASUREMENT – MENTAL HEALTH OVERVIEW: FINAL REPORT

Name	Purpose	Multi-Dimensional?	Reliability	Validity	Change?	Time	Acceptability	Who?	Electronic?	Interpretation	Cost
MMPI (-1, -2 adults and -A adolescent)	<p>?To aide in clinical diagnosis</p> <p>?Assessment of general psychopathology in adults</p> <p>?Screening for wide variety of mental disorder symptomatology</p>		<p>?Cronbach's alpha for basic scales: .34-.88</p>	<p>?Research base from 60 years, with > 10,000 citations supports validity across wide range of applications (psychiatric, general medical, forensic, vocational)</p> <p>?Better evidence for convergent than for discriminative validity</p> <p>?Not all newer supplementary scales have been adequately validated</p> <p>?Sensitivity and specificity</p>	<p>Nor reported by APA</p>	<p>1-1.5 hours for self-report (booklet or audio tape)</p>	<p>Probably too long for routine use, booklet requires minimal reading level of 8th grade</p>	<p>Self report (booklet or audio tape)</p>	<p>Yes</p>	<p>8 basic syndrome scales: hypochondriasis, depression, hysteria, psychopathic deviate, paranoia, psychasthenia, schizophrenia, mania, additional supplementary scales (inc. PTSD) can also be scored raw and T-scores interpretation varies according to context</p>	<p>Hand scoring starter kit = \$350</p>

OUTCOME MEASUREMENT – MENTAL HEALTH OVERVIEW: FINAL REPORT

Name	Purpose	Multi-Dimensional?	Reliability	Validity	Change?	Time	Acceptability	Who?	Electronic?	Interpretation	Cost
PFI (Feragne, 1983)	<p>?To meet the need for a comprehensive, general-purpose and multi-dimensional outcome measure.</p> <p>?Designed as part of a study of psychiatric inpatients</p> <p>?To consolidate in a single instrument the equivalent of a comprehensive "core battery"</p>	Yes	<p>?Cronbach's alpha at admission : .75-.88 (except spouse and house mate role scales)</p> <p>?Immediate post-discharge , alpha = .55-.96 for all scales</p>	good	yes	?45 minute by interviewer (recommended due to possible problems with self-completion by psychotic patients	Not specifically addressed by Feragne (1983)	<p>?Self-completed or structured interview</p> <p>?No special training required</p> <p>?Amount of any training would be function of interviewer experience with population</p> <p>?2 hours didactic training + 5-6b observed interviews provides adequate training</p>	Not noted		
VA OPCS Technology Assessment Program				www.va.gov/vetap							23

OUTCOME MEASUREMENT – MENTAL HEALTH OVERVIEW: FINAL REPORT

Name	Purpose	Multi-Dimensional?	Reliability	Validity	Change?	Time	Acceptability	Who?	Electronic?	Interpretation	Cost
SCL-90-R	Intents: ?Quick screening ?Outcome of psychopathology ?Quantifying current psychopathology	Yes	Good	Better convergent than divergent validity	“Some evidence”	12-20 minutes, after brief intro & instruction to assure validity	Widely used as screening instrument for global psychological distress, multidimensional symptom profile	Self-administered	Yes (\$89)	Raw scores and T-values for 9 dimensions (somatization, obsessive-compulsive, interpersonal sensitivity, depression, anxiety, phobic-anxiety, hostility, paranoid ideation, psychotics 3 global indices, global severity index	Copyrighted, hand scoring starter kit is \$104.

OUTCOME MEASUREMENT – MENTAL HEALTH OVERVIEW: FINAL REPORT

Name	Purpose	Multi-Dimensional?	Reliability	Validity	Change?	Time	Acceptability	Who?	Electronic?	Interpretation	Cost
TAG under development and testing (Slade, 2000)	Brief assessment of severity of mental illness		Under investigation (Slade 2000)	High, by virtue of development process (search and Delphi groups)	Not explicitly noted by Slade (2000)	1-3 minutes	Probably good, as clinicians participated in development	Clinician	Not noted by Slade (2000)	Total number of judgments of severity level for each of 7 domains	Public domain

Abbreviations: APA, American Psychiatric Association (*Handbook of Psychiatric Measures* (2000))

- BASIS-32, Behavior and Symptom Identification Scale
- BPRS, Brief Psychiatric Rating Scale
- BASIS-32, Behavior and Symptom Identification Scale
- BSI, Brief Symptom inventory
- CAN, Camberwell Assessment of Need (Slade 1999b; Phelan, 1995)
- CGI, Clinical global impressions scale
- COMPASS-OP, Compass out-Patient
- COOP, Dartmouth COOP Functional Assessment Charts
- DUKE, Duke Health Profile
- GAF, Global Assessment of Function
- GHQ, General Health Questionnaire
- HoNOS, Health of the Nation Outcomes Scales
- PFI, Psychosocial Functioning Inventory (Feragne, 1983)
- JCAHO, Joint Commission on Accreditation of Healthcare Organizations
- MCAS, Multinomial Community Ability Scale
- MHI, Mental Health Inventory
- MHQ, Middlesex Hospital Questionnaire
- MOT, Medical Outcomes Trust
- PFI, Psychosocial Functioning Inventory (Feragne, 1983)
- SCL-90-R, Symptom Checklist–90, revised

SIP, Sickness Impact Profile
SPSS, Statistical Package for the Social Sciences
TAG, Threshold Assessment Grid (Slade, 2000)

TABLE 2. Measures Evaluation Matrix Summary: Global Mental Health Status, Psychiatric Symptoms, Functioning, Disabilities Measures (Specific To Mental Health, Not General Health Or Emotional Health Secondary To A General Medical Condition)

Note: Unless otherwise noted, the information in this table was obtained from The American Psychiatric Association’s *Handbook of Psychiatric Measures* (2000). Shaded rows indicate that five standardized instruments appear to meet most of VHA’s needs.

Criteria for VA use of standardized measures of mental health care outcome:

- | | |
|---|---|
| <ol style="list-style-type: none"> 1. Original purpose congruent with intended VA use for quality of care tracking and reporting, documenting effective treatment of veteran mental health patients treated in a variety of settings (measures combining functional status and symptom severity deemed the best fit = X); 2. Multi-dimensional; 3. Acceptable psychometrics; 4. Sensitive to change; | <ol style="list-style-type: none"> 5. Feasible for routine use (i.e., brief, imposing minimal administrative, clinical, and respondent burdens); 6. Electronic data entry and/or analysis available; 7. Readily interpretable by non-professionals (i.e., summary score of generally understood construct not requiring extensive background information for interpretation and immediate understanding); 8. Low cost (i.e., measure in the public domain or available for distribution and use at minimal cost to VA). |
|---|---|

Measure	Criterion key (from list above)							
	1	2	3	4	5	6	7	8
GAF	X		X	X	?	X	X	X
HoNOS	<u>X</u>		X	X	X	X	X	X
BASIS-32	<u>X</u>		X	X	X	X	X	X
BPRS	X		X	X	X	?	X	X
COMPASS-OP	X		X	X	X	X	X	?
BSI	X		X	?	X	X		
CAN	X	X	X	?	?		X	X
CGI	X		?	X	X		X	X
GHQ								
MCAS			X	?	X		X	X
MHI			X					

OUTCOME MEASUREMENT – MENTAL HEALTH OVERVIEW: FINAL REPORT

MHQ	X		?	X	X		X	?
MMPI			X	?		X		
PFI	X	X	X	X			X	?
SCL-90-R	X		?	X	X	X		
TAG	X		?	?	X		X	X

Abbreviations: See Table 1

VIII. REFERENCES

- Andrews G, Peters L, Teeson M. (The Australian Health Outcomes Collaboration). Measurement of consumer outcome in mental health. 1994. <http://www.crufad.com/books/consumerout.htm>.
- Barry MM, Zissi A. Quality of life as an outcome measure in evaluating mental health services: a review of the empirical evidence. *Social Psychiatry and Psychiatric Epidemiology*, 1997; 32: 38-47. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=9029986.
- Bebbington P, Brugha T, Hill T, Marsden L, Window S. Validation of the Health of the Nation Outcome Scales. *British Journal of Psychiatry*, 1999; 174: 389-394.
- Blais MA, Lenderking WR, Baer L, deLorell A, Peets K, Leahy L, et al. Development and initial validation of a brief mental health outcome measure. *Journal of Personality Assessment*, 1999; 73: 359-373. <http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?db=m&form=6&dopt=r&uid=10689649>.
- Brugha TS, Lindsay F. Quality of mental health service care: the forgotten pathway from process to outcome. *Social Psychiatry and Psychiatric Epidemiology*, 1996; 31: 89-98.
- Clifford PI, Orbach G, Hobbins J, Pavyer J. Measuring disability and outcomes in routine practice with the FACE Core Assessment. *Bulletin of the Menninger Clinic*, 1999; 63: 332-345.
- Coffey M, Jones SH, Thornicroft G. A brief mental health outcome scale: relationships between scale scores and diagnostic/sociodemographic variables in the long-term mentally ill. *Journal of Psychiatric and Mental Health Nursing*, 1996; 3: 89-93. <http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?db=m&form=6&dopt=r&uid=8716799>.
- Dickerson FB. Assessing clinical outcomes: the community functioning of persons with serious mental illness. *Psychiatric Services*, 1997; 48: 897-902. <http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?db=m&form=6&dopt=r&uid=9219297>.
- Ehrlichman M. Personal communication. Fax and Email To: Flynn K; May 31, 2001, 2001.
- Eisen SV, Dill DL, Grob MC. Reliability and validity of a brief patient-report instrument for psychiatric outcome evaluation. *Hospital and Community Psychiatry*, 1994; 45: 242-247. <http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?db=m&form=6&dopt=r&uid=8188195>.
- Eisen SV, Wilcox M, Leff HS, Schaefer E, Culhane MA. Assessing behavioral health outcomes in outpatient programs: reliability and validity of the BASIS-32. *Journal of*

OUTCOME MEASUREMENT – MENTAL HEALTH OVERVIEW: FINAL REPORT

Behavioral Health Services Research, 1999; 26: 5-17.

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=10069137.

Feragne MA, Longabaugh R, Stevenson JF. The psychosocial functioning inventory. Evaluation and the Health Professions, 1983; 6: 25-48.

Gallagher J, Teesson M. Measuring disability, need and outcome in Australian community mental health services. Australian and New Zealand Journal of Psychiatry, 2000; 34: 850-855. <http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?db=m&form=6&dopt=r&uid=11037373>.

Greenley JR, Greenberg JS, Brown R. Measuring quality of life: a new and practical survey instrument. Social Work, 1997; 42: 244-254. <http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?db=m&form=6&dopt=r&uid=9153093>.

Hafkenscheid A. Psychometric measures of individual change: an empirical comparison with the Brief Psychiatric Rating Scale (BPRS). Acta Psychiatrica Scandinavica, 2000; 101: 235-242.

Holohan TV, Lehmann LS. Personal communication. Telephone To: Flynn K; August 7, 2001.

Huxley P. Outcomes management in mental health: A brief review. Journal of Mental Health, 1998; 7: 273-283.

Huxley P, Reilly S, Gater R, Robinshaw E, Harrison J, Mohamad H, et al. Matching resources to care: the acceptability, validity and inter-rater reliability of a new instrument to assess severe mental illness (MARC-1). Social Psychiatry and Psychiatric Epidemiology, 2000; 35: 312-317.

Jansen DP. (VHA Mental Health Strategic Health Care Group). Need for mental health services in the veterans Health Administration : a report prepared for the Under Secretary for health, Veterans Health Administration, U.S. Department of Veterans Affairs. June 10, 2001. Available: June 10.

Jenkins R. Towards a system of outcome indicators for mental health care. British Journal of Psychiatry, 1990; 157: 500-514.

Kamis-Gould E, Waizer J. National data standards for mental health management and performance indicators. The Psychiatric Hospital, 1992; 23: 23-28.

Lehmann LS. Personal communication. Electronic Mail To: Flynn K; June, 2001.

Lohr KN. Outcome measurement: concepts and questions. Inquiry, 1988; 25: 37-50. <http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?db=m&form=6&dopt=r&uid=2966125>.

OUTCOME MEASUREMENT – MENTAL HEALTH OVERVIEW: FINAL REPORT

Lohr KN. Applications of health status assessment measures in clinical practice. Overview of the third conference on advances in health status assessment. *Medical Care*, 1992; 30: MS1-14. <http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?db=m&form=6&dopt=r&uid=1583924>.

Lohr KN, Aaronson NK, Alonso J, Burnam MA, Patrick DL, Perrin EB, et al. Evaluating quality-of-life and health status instruments: development of scientific review criteria. *Clinical Therapeutics*, 1996; 18: 979-992. <http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?db=m&form=6&dopt=r&uid=8930436>.

McClelland R, Trimble P, Fox ML, Stevenson MR, Bell B. Validation of an outcome scale for use in adult psychiatric practice. *Quality in Health Care*, 2000; 9: 98-105. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=11067258.

McCullough L. Standard and individualized psychotherapy outcome measures: A core battery. In: Miller NE, Luborsky L, Doherty JP, Eds. *Psychodynamic treatment research: A handbook for clinical practice*. New York: Basic Books, 1993: 469-496.

Moos RH, McCoy L, Moos BS. Global Assessment of Functioning (GAF) ratings: Determinants and role s predictors of one-year treatment outcomes. *Journal of Clinical Psychology*, 2000; 56: 449-461.

Newman FL, Hunter RH, Irving D. Simple measures of progress and outcome in the evaluation of mental health services. *Evaluation and Program Planning*, 1987; 10: 209-218.

Orrell M, Yard P, Handysides J, Schapira R. Validity and reliability of the Health of the Nation Outcome Scales in psychiatric patients in the community. *British Journal of Psychiatry*, 1999; 174: 409-412.

Pallak MS. National outcomes management survey: summary report. *Behavioral Healthcare Tomorrow*, 1994; 3: 63-65, 69.

Phelan M, Slade M, Thornicroft G, Dunn G, Holloway F, Wykes T, et al. The Camberwell Assessment of Need: the validity and reliability of an instrument to assess the needs of people with severe mental illness. *British Journal of Psychiatry*, 1995; 167: 589-595. <http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?db=m&form=6&dopt=r&uid=8564313>.

Piersma HL, Boes JL. The GAF and psychiatric outcome: a descriptive report. *Community Mental Health Journal*, 1997; 33: 35-41. <http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?db=m&form=6&dopt=r&uid=9061261>.

Powell JA, Geddes JR. Evidence-based psychiatry [Letter]. *British Journal of Psychiatry*, 1997; 171: 586-587. <http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?db=m&form=6&dopt=r&uid=9519112>.

Reus VI. Mental Disorders. In: Fauci AS, Braunwald E, Isselbacher KJ, Wilson JD, Martin JB, Kasper DL, et al., Eds. *Harrison's Principles of Internal Medicine*. New York: McGraw-Hill, 1998: 2485-2503.

Rosenheck R, Van Stone W. Evaluating VA outcomes using the Global Assessment of Functioning (GAF). Unpublished Manuscript. July 16, 2001.

Ruggeri M, Leese M, Thornicroft G, Bisoffi G, Tansella M. Definition and prevalence of severe and persistent mental illness. *British Journal of Psychiatry*, 2000; 177: 149-155.

Sederer LI, Dickey B, Eisen SV. Assessing outcomes in clinical practice. *Psychiatric Quarterly*, 1997; 68: 311-325. <http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?db=m&form=6&dopt=r&uid=9355132>.

Sederer LI, Hermann R, Dickey B. The imperative of outcome assessment in psychiatry. *American Journal of Medical Quality*, 1995; 10: 127-132.

Sharma VK, Wilkinson G, Fear S. Health of the Nation Outcome Scales: a case study in general psychiatry [see comments]. *British Journal of Psychiatry*, 1999; 174: 395-398.

Slade M. Personal communication. Electronic mail To: Flynn K; June 26, 2001.

Slade M, Beck A, Bindman J, Thornicroft G, Wright S. Routine clinical outcome measures for patients with severe mental illness: CANSAS and HoNOS. *British Journal of Psychiatry*, 1999; 174: 404-408. <http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?db=m&form=6&dopt=r&uid=10616605>.

Slade M, Thornicroft G, Glover G. The feasibility of routine outcome measures in mental health. *Social Psychiatry and Psychiatric Epidemiology*, 1999; 34: 243-249.

Slade M, Thornicroft G, Loftus L, Phelan M, Wykes T. CAN: Camberwell Assessment of Need. A comprehensive needs assessment tool for people with severe mental illness. London: Gaskell (Royal College of Psychiatrists), 1999.

Slutsky J. Guideline dissemination and changing behavior : getting guideleines into the hands of policy makers and clinicians. 2001.

Smith GR. Considerations in choosing, using, and interpreting measures for health care systems. In: Rush AJ, Ed. *Handbook of Psychiatric Measures / Task Force for the Handbook of Psychiatric Measures*. Washington, DC: American Psychiatric Association, 2000.

Stedman T, Yellowlees P, G. M, Clarke R, Drake S. (ACT: Department of health and Family Services). Measuring consumer outcomes in mental health: field testing of selected measures of consumer outcome in mental health. October, 1997. Available: October. <http://www.mentalhealth.gov.au/mhinfo/ccf/pdf/outcomes.pdf>.

OUTCOME MEASUREMENT – MENTAL HEALTH OVERVIEW: FINAL REPORT

Tancredi LR. Defining, measuring, and evaluating quality of care. In: Sloan FA, Blumstein JF, Perrin JM, Eds. *Cost, Quality, and Access in Health Care: new roles for health planning in a Competitive Environment*. San Francisco: Jossey-Bass, 1998.

Trauer T. Issues in the assessment of outcome in mental health. *Australian New Zealand Journal of Psychiatry*, 1998; 32: 337-343.

Trauer T, Callaly T, Hantz P, Little J, Shields R, Smith J. Health of the Nation Outcome Scales. Results of the Victorian field trial. *British Journal of Psychiatry*, 1999; 174: 380-388.

Zarin DA. Considerations in choosing, using, and interpreting a measure for a particular clinical context. In: Rush AJ, Ed. *Handbook of Psychiatric Measures*. Washington, DC: American Psychiatric Association, 2000.



**T e c h n o l o g y
A s s e s s m e n t
P r o g r a m**

Office of Patient Care Services

**VA Technology Assessment Program
Office of Patient Care Services (11T)**

OUTCOME MEASUREMENT – MENTAL HEALTH OVERVIEW: FINAL REPORT

**VA Boston Healthcare System
150 South Huntington Avenue
Boston, MA 02130**

Tel: 617.278.4469 Fax: 617.264.6587

vatap@med.va.gov

<http://www.va.gov/vatap> <http://vaww.va.gov/vatap>

Report released: July 2002



**T e c h n o l o g y
A s s e s s m e n t
P r o g r a m**

Office of Patient Care Services

**OUTCOME MEASUREMENT
IN VHA MENTAL HEALTH SERVICES**

**An Overview and Series of
Diagnosis Specific Short Reports**

APPENDIX

Author: Karen Flynn, MS, DDS
Manager, Technology Assessment Program

Contributors: Elizabeth Adams, RRT, MPH
Research Analyst

Elaine Alligood, MLS
Information Specialist

Report Released: July 2002

APPENDIX

1. INTRODUCTION

The intent of this Appendix is to provide interested readers with additional background information on the development and psychometric characteristics of health measurement instruments to that supplied in the “Overview” and diagnostic specific sections of the report.

Borrowing its perspective from one of the primary sources for the Appendix (Blacker and Endicott, 2000), the focus is conceptual rather than technical. It will thus guide the reader through general background concepts to their application in a set of potential selection criteria relevant to VHA as it chooses standardized outcome measurement instruments for its mental health services. Readers needing additional technical information should consult corresponding references suggested by Blacker and Endicott, or those at the end of this Appendix.

II. TYPES OF HEALTH MEASUREMENTS

McDowell and Newell (1996) discuss the types of health status measurement. Terms from McDowell and Newell are defined below; they appear in the names of standardized instruments referenced in these reports, and thus make a useful beginning to this appendix: Unless otherwise noted, definitions are those given by McDowell and Newell (1996).

Clinical interviews measure the health status of individuals.

Diagnostic Indices include physiologic measurements such as those of blood pressure, thyroid stimulating hormone, serum creatinine, or hemoglobin.

Discriminative indices, such as IQ tests, are designed to classify people when no external criterion exists.

Evaluative indices measure change over time.

Health index is a measurement that yields a single summary score.

Health profile: scores on the different dimensions of an instrument are presented separately.

Macro health indices measure group health status change.

Outcomes management refers to outcomes measurement in continuous quality improvement programs.

Prognostic measures or predictive classify people according to some criterion, which may exist now (diagnosis) or in the future (prognosis), such as those predicting the ability to live independently in the community following rehabilitation.

Rating scale is used by an expert, usually a clinician, to assess defined aspects of health, but the precise questions can vary from rater to rater and subject to subject. This approach is based on the belief that a flexible approach to measuring different people will capture more modulated, but still equivalent information, while compensating for differences in the use of language.

The **repertory grid** classifies people's thoughts on two dimensions: elements or topics thought about; and constructs or qualities used to define and think about elements. Measurements of this type are beginning to be used in quality of life measurement.

Questionnaires are often self-completed, and like **interview schedules** contain pre-set questions; interviewers are trained not to alter question wording. Questionnaires and interview schedules thus standardize assessments across individuals (assuming that a standard set of measurement dimensions or scales is equivalently relevant to every person being measured and that scoring procedures should remain constant).

Standardized measures are those for which psychometric properties (validity, reliability, sensitivity to change) have been demonstrated (Slade, 1999).

Survey instruments measure the health status of groups.

III. THE QUALITY OF A MEASUREMENT AS INDICATED BY ITS PSYCHOMETRIC PROPERTIES: VALIDITY, RELIABILITY, FEASIBILITY

Blacker and Endicott (2000) remind us of the original meaning of **psychometrics**: “mind measuring”, and of the term's present general use to indicate performance characteristics of many types of measures.

*“The psychometric concepts of reliability and validity are fundamental to critical review of measurement instruments in psychiatry. Reliability is the dependability or reproducibility of measurement and is typically assessed in three ways: **inter-rater reliability**, reflecting the likelihood that two or more raters will agree on their judgments; **test-retest reliability**, referring to the stability of measurements at different points in time; and **internal consistency**, referring to the homogeneity of certain scales, whether all items measure the same construct. Once reliability has been established, then we know that measurement is reproducible. But we do not know what it measures.”*Wetzler (1989)

McDowell and Newell (1996) provide further background to the field of psychometrics: the ability of humans to act as accurate measuring instruments has been under investigation since the middle of the nineteenth century. Psychometrics grew from psychophysics, or investigations into the ways in which people perceive and make

judgments about physical phenomena, initially those for which objective physical measurements exist.

However, the social and health sciences often are concerned with subjective judgments regarding phenomena for which there are no objective measurements. The field of psychometrics studies the measurement of qualities for which there is no absolute physical scale, and thus is central to any discussion of health status measurement, as in this series of VATAP reports.

A. Validity

Validity is the property of a measurement instrument concerning whether it measures what it purports to measure, or the accuracy of its representation of the “state of nature”. The same approach underlies the notions of sensitivity and specificity of diagnostic tests.

While **reliability** is an empirical question (see discussion below), validity is partly theoretical; for many constructs measured in psychiatry there is no absolute truth. However, since some measurement instruments yield more useful and meaningful data than others, assessment of validity is valuable.

Quantitative validity assessments are possible to the extent that there is agreement on a gold standard or criterion of accuracy against which an instrument can be tested. Testing to see if the instrument correlates as expected to the construct under study is another approach to quantifying validity (Blacker and Endicott, 2000).

Face validity is that aspect of overall validity describing whether an instrument’s items appear to assess the construct of interest.

Content validity describes the adequacy of coverage of relevant domains. The assessment of this aspect of validity is fundamentally qualitative and depends on a careful inspection of each item by someone who thoroughly understands the intent of the instrument. Users must feel comfortable with specific questions and their wording. Developers and users should have conceptualized the construct or constructs of interest in the same way. (Blacker and Endicott, 2000).

Assessing validity against a gold (or criterion) standard

Criterion, predictive, or concurrent validity deals with an instrument’s agreement with a gold standard or criterion of accuracy. In cases where an instrument results in categorical classification (or a continuous scale with a cut-point), criterion validity is evaluated in ways analogous to diagnostic test evaluations (Flynn, 1996).

Assessing validity without an adequate gold standard

Construct validity assesses whether an instrument correlates as expected with external validators (which may include other, well-established instruments), and may be useful either when an adequate gold standard is not available, or when additional validity data are desired. **External validators** are attributes that bear a well-characterized relationship to the construct of interest but are not measured directly by the instrument being validated. Instruments that purport to measure a given construct are validated on the basis of their ability to identify individuals with the expected attribute. External validators are less clearly related to direct measurement of the construct of interest than is a gold standard, although the two approaches can overlap (Blacker and Endicott, 2000).

Discriminant validity tests an instrument's ability to discriminate between populations that are expected to differ on the construct of interest. A study of discriminant validity is more clearly relevant if it includes the types of cases encountered in clinical practice, i.e., can discriminate "at the margins" of the distributions of characteristics within clear cut and mutually exclusive diagnostic classifications (Blacker and Endicott).

Factor analysis examines whether the interrelationships among items demonstrate the expected structure for the construct. With factor analysis, one can also include items from other instruments that measure similar and dissimilar constructs. Results from factor analysis depend on the sample.

Convergent validity indicates the statistical association of items from the construct under study with items from measures of similar constructs and with different factors from measures of different constructs (**divergent validity**) (Blacker and Endicott).

"Assessing validity in areas in which there are few established measures and for which a gold standard or criterion of accuracy cannot be established is difficult. The assessment of the validity of the measure is essentially a joint measure of the validity of the measure and the validity of the construct itself. For this reason, it is most problematic when it is most needed – for measures of newer, less validated constructs. Moreover, the various external validators may not all yield the same answer about the optimal measure or optimal definition of a construct. Nonetheless, by triangulating between a better definition of the construct, better ways to measure it, and better exploration of how it operates in clinical practice and research, the field moves to greater validity over time." Blacker and Endicott (2000).

Validity can also be demonstrated by **sensitivity to change** (i.e., improvement with efficacious treatment or decrement with disease progression). To be optimally useful, sensitivity to change data must reflect amounts of change that are clinically relevant and the range of scores in which change is detected should be relevant in a given setting (Blacker and Endicott, 2000).

According to McDowell and Newell (1996), the common definition of validity above is insufficient, since valuable interpretations of a test beyond its original intent may be

found. McDowell and Newell (1996) propose a more general definition: “*validity describes the range of interpretations that can appropriately be placed on a measurement score: What do the results mean? What can we conclude about the person who produced particular scores on the test? By focussing attention on the breadth of a measure, this approach also reflects its sensitivity. The shift in definition is significant, for validity is no longer a property of the measurement, but rather of the interpretation we place on the results... Sloppiness in defining the precise purpose of a measurement... can best be avoided by closely linking the validation process to a conceptual expression of the aims of the measurement and also linking the concept with other, related concepts to indicate other alternative possible interpretations of scores.*”

The choice of methods to measure validity depends on its purpose (i.e., screening test or outcome measurement). Most validation studies begin by referring to **content validity** (comprehensiveness, or how adequately the sampling of questions reflects the aims of the index that were specified in the conceptual definition of its scope (for instance: Are all of the items relevant to the concept? Are all aspects of the concept covered?)) **Content validity** is seldom formally tested. Rather, the face validity or clinical credibility of a measure is commonly inferred from expert review of its clarity and completeness. Following content validation, formal statistical procedures are used to assess validity. In the simplest case, there is already an accepted way to measure the concept in question, a criterion or gold standard against which the new instrument may be compared. The new and established tests are applied to a suitable sample of people, and the results are compared using an appropriate indicator of agreement (often a correlation coefficient).

B. Reliability

Reliability is “*The consistency or precision with which a measure can discriminate one subject from another*”, in other words consistent and repeatable even if performed by different raters at different times or under different conditions. Reliability has three standard forms: **internal consistency**, **inter-rater reliability**, and **test-retest reliability**. (Blacker and Endicott, 2000).

*“The level of reliability indicates the highest degree of validity possible. Outcomes measured unreliably can never be valid, and if determining the reliability of an outcome indicator is difficult, determining its validity will be even more problematic.”*Lohr (1988)

Constructs assessed in psychiatric research and practice include: diagnosis, signs and symptoms, impairment, patient and family functioning, met or unmet needs, and quality of life. Constructs may be measured using **categorical** (qualitative, with a finite number of categorical options) or **continuous** (quantitative, along a continuum of intensity, frequency, or severity) classifications and data. Each construct assessed in an outcome measure may have single or multiple **domains or sub-domains**, which in turn may be related or independent (Blacker and Endicott, 2000).

Internal consistency is a measure of agreement among the individual components of an instrument. Evaluation for internal consistency views each item as a single measurement of the underlying construct; coherence of individual items thus suggests that they all, individually and collectively, measure the same thing. Internal consistency is higher to the extent that all items are measuring the same dimension, and is usually reported with **Cronbach's alpha** (the degree to which items co-vary with each other).

Internal consistency of an instrument, of course, depends on the consistency of the construct that it was developed to measure, and measures designed to assess a multidimensional construct may be better judged by the internal consistency of each relevant sub-scale. Internal consistency can be readily measured from a single administration of the instrument, and is therefore well suited to quick initial assessments of performance.

Item-total correlation (of each item with the total score) is an aspect of internal consistency that may be reported as an indicator of **reliability**.

Inter-rater reliability measures the agreement between two or more observers of the same group of subjects, using the same information.

Test-retest reliability measures the agreement between evaluations at two points in time and, unlike inter-rater reliability, is suitable for self-report measures, to the extent that a subject's true condition remains stable during the time interval. To avoid subject-associated change related to time, test-retest reliability studies need to select a time interval short enough that little or no change occurs, but long enough that respondents are unlikely to remember their answers on the initial administration. In practice, no interval satisfies these requirements; for conditions that fluctuate over brief periods; compromises are inevitable.

Both inter-rater and test-retest reliability are reported using the same statistic: for categorical data, the **kappa coefficient** (a measure of agreement corrected for chance agreement; kappa = 1 indicates perfect agreement; kappa = 0 indicates only chance agreement; kappa >0.8 may be considered excellent; 0.7-0.8, good; 0.5-0.7 fair; and <0.5 poor).

For ordinal scales, the weighted kappa adjusts the level of disagreement for the distance between the two ratings. Continuous scales use intra-class correlation coefficients.

The need for reliability varies with the setting in which an instrument will be used. Where small differences are clinically important, outcome instruments will require high reliability; for other purposes, good to fair reliability may be adequate. Extensive rater training and optimal testing conditions in published reliability studies tend to overestimate the reliability that can be obtained in routine clinical practice. Reliability also depends on the nature of the sample in which the measure was tested. A large proportion of individuals near a diagnostic threshold or resistant to instrument administration will diminish observed reliability. Finally, since greater effort is required to make fine distinctions, reliability tends to be higher in samples with high variability,

making it risky to generalize a reliability estimate from a heterogeneous population to a homogeneous one (Blacker and Endicott, 2000).

C. Feasibility

Slade (1999) proposes that another property critical to the use of routine outcome measures in mental health care be classified as a psychometric property: Slade defines **feasibility** as an instrument's suitability for routine, sustained, and meaningful use in typical clinical settings. In the rationale for feasibility, Slade also notes the research context for the development of most instruments. That context tends to lack a focus on feasibility, which Slade further defines by six characteristics: brevity (including at a first use); relevance (avoiding jargon and allowing flexibility in how questions are worded enhances relevance to the patient; from the staff perspective, relevance indicates a concurrence between measurement results and clinical judgment; acceptability (what is assessed, the purpose, method, and language used in the assessment; availability (ease of obtaining and/or photocopying the instrument; and value (staff see the benefits of a standardized assessment as outweighing the costs, and feed-back to staff occurs in a constructive way).

Slade suggests that feasibility can only be demonstrated in the context of a particular use and setting; further, new methods of instrument development are needed and feasibility concerns should be recognized from the beginning of the development process. One readily notes that feasibility is likely to be setting-specific, as well. For example, feasibility for VHA use would most likely require demonstration by pilot testing a new instrument within VHA mental health programs, at the facilities where it ultimately would be routinely used.

IV. PREREQUISITES FOR HEALTH INDICES

Eisen (1991) notes that precise definition of the goals of an outcome evaluation in mental health care should precede instrument choice. Among the goal-associated decisions are: which treatment effects are to be measured, and whose perspective (administrators, treaters, patients, or their collaterals) is to be emphasized.

Sackett (1977) lists essential characteristics of health indices, when these indices will be used for comparisons in the evaluation of health care programs, and when indices need to be credible to both clinicians and administrators:

- ?? *Positive orientation*, i.e. document the presence of health, rather than being confined to cataloging symptoms, illnesses, or catastrophes;
- ?? *Comprehensiveness*, i.e., encompassing several domains of health (social, emotional, physical);
- ?? *General applicability* to people in a range of treatment settings, and to those not under active treatment;
- ?? *Sensitivity* to important changes in health status or function;

- ?? *Simplicity, acceptability, and cost*, analogous to Slade’s feasibility, with the addition of ratings being made or scored by non-clinicians;
- ?? *Precision*, i.e., with high reproducibility of individual and group measurements at short intervals (e.g., *reliability*);
- ?? *Amenability to index construction*, i.e., measurements would permit rapid combination into composite indices (thus, qualitative responses or measurements requiring content analysis should be avoided).

Owen and Thrush (2000) note the recent shift from process measurement to patient outcomes in quality assessment. These authors contributed a chapter on instruments suited to health care system evaluation to the *APA Handbook of Psychiatric Measures*. In this chapter, they note that both generic and disease-specific outcomes assessment instruments collect patient-specific data that can be aggregated to examine outcomes at the system level.

V. CRITERIA FOR A VHA MENTAL HEALTH CARE OUTCOME MEASURE

Owen and Thrush (2000) further note that any symptom that is sensitive to change with treatment can function as an outcome measure. Thus, generic and disease specific measures other than those that focus their chapter can also serve, if aggregated across patients, as system evaluation measures.

According to Owen and Thrush, measures that also capture information about patient characteristics and processes of care allow for case-mix adjustment and therefore are uniquely suited to practitioner and system evaluations. However, VHA’s internal data systems for patient and process characteristics may ease restrictions relative to instruments that also capture these data on its choice of instruments for system-wide use in mental health care. Further, Owen and Thrush include only one generic (COMPASS-OP for outpatients) and one schizophrenia-specific instrument relevant to this review (Schizophrenia Outcomes Module), both of which do not meet criteria for VHA use on other grounds (see Tables 1 and 2 of the Overview and Table 1 of the Schizophrenia section).

Smith (2000) suggests a general framework within which to organize the analysis of measures for use in health care systems. This author categorizes as “Type B” those measures using individual patients as data sources to provide information about systems of care. Such measures are usually “Type A” (or an individual patients provides information for that particular patient) measures used in data collection efforts representative of the system of care. For VHA mental health services, system representation would likely mean sampling for instrument administration from the overall mental health patient population in ways that adequately represent the demographics, diagnoses, severities of illness, levels of care, and relevant co-morbidities seen in that population.

VHA’s HSR&D Service could provide survey research experts to assist in designing a sampling strategy. Patient-level data can be aggregated to reflect a system of care, if

sampling is systematic and can be demonstrated to reflect the characteristics of patients in the system as a whole. Smith (2000) reports that, while mental health clinicians have long relied on general symptom inventories, general and disease-specific approaches to the assessment of outcomes are now frequently combined. The optimal approach for a system is determined by the degree of detail desired (disease-specific assessments provide more in-depth information within a diagnostic group), and by the extent to which comparisons of outcomes across diagnostic groups would be useful.

Newman (1987) advocates strongly for the use of both global and carefully selected specific measures. Newman further provides examples of ways in which the data from such combinations can be used for both clinical and administrative purposes.

The Medical Outcomes Trust is a non-profit agency devoted to the collection, quality assessment, and dissemination of standardized outcomes assessment instruments (Lohr, 1996). The Scientific Advisory Committee of the Trust developed a set of attributes for the review and evaluation of instruments, the first six of which are relevant to VHA:

1. Conceptual and measurement model:

- ?? What is the basis for combining items into scales?
- ?? What descriptive statistics for the scales can be provided? Do the scales demonstrate adequate variability?
- ?? What evidence describes or supports the intended level of measurement? (ordinal, interval, or ratio scales)
- ?? Are the procedures for deriving scale scores from raw scores adequately justified and specified?

2. Reliability: the degree to which the instrument is free from random error:

- ?? Internal consistency
- ?? Reproducibility

3. Validity: the degree to which the instrument measures what it purports to measure:

- ?? Content
- ?? Construct
- ?? Criterion

4. Responsiveness: the instrument's ability to detect changes in outcome that matter to persons with the health condition, to their significant others, or to their health care providers (sensitivity to change).

5. Interpretability: the degree to which one can assign qualitative meaning (clinical or commonly understood connotations) to quantitative scores, by means of various types of information to assist in interpretation:

- ?? Comparative data on the distribution of scores from other groups, including the general public;
- ?? The relationship of scores to various events (losing a job, graduating from college, or needing institutional care);

?? Clear descriptions of the comparison populations and the means by which relevant data were amassed, interpreted and displayed.

6. Respondent and administrative burden: the time, energy, financial resources, personnel, needed to administer the instrument.

As of September 1996, six instruments met review criteria and were included in the Trust's library. The Trust catalog dated 2000 contains 12 generic measures instruments, and another 12 condition-specific ones, although none of the latter are specific to serious mental illnesses.

Medical Outcomes Trust-approved instruments will be indicated in further sections of this review. VATAP assigned a high degree of credibility to Trust approval, as the criteria used are the most rigorous currently on record.

VI. SELECTION CRITERIA FOR STANDARDIZED MENTAL HEALTH CARE OUTCOMES MEASURES FOR VHA

Final selection of instruments for VHA use will depend on clinician and administrator preferences for certain constructs and on intended use. Decisions will involve construct preferences: health, disease, function, quality of life, patient satisfaction, met or unmet need for services, patient or clinician perspective. Some of these decisions will necessarily overlap with intended use and practical issues of administration, data analysis, and reporting.

While some of these constructs do not necessarily adhere to strict application of Lohr's 1988, "classical" definition of outcomes, such arguments may be largely academic relative to other overriding considerations such as VHA stakeholder interests, or VHA's mission. VHA will undoubtedly select for routine use those instruments that best meet the widest range of preferences, needs, and practical constraints.

For example, VHA may choose to demonstrate the quality of its mental health care using generic instruments or diagnosis specific ones. The choice will be determined by the uses to which the resulting information will be put. Global instruments, applicable across a variety of diagnoses and treatment settings, would allow greater flexibility in comparing outcomes across diagnoses, programs, and facilities within the system, or with other systems (i.e., for "benchmarking") in which the instruments are also used.

VATAP took the well-reasoned advice of Newman (1987) in providing information on both global and specific instruments in this series of reports. Finally, other large national health care systems engaged in similar mental health care quality reporting endeavors (Slade, 2001; Stedman, 1997) have focused on global measures when making recommendations for routine use in mental health services.

VHA Mental Health Services' overall emphasis on returning patients to the highest degree of function possible would be well served by the use of generic functional status

instruments. Generic instruments also would provide more straightforward summary information on the care of patients within general classifications, such as serious mental illness, dual diagnoses, or the elderly with mental illness.

Diagnosis specific instruments, on the other hand, would allow for greater detail in comparisons within groups of patients with the diagnoses, or in reporting on outcomes in these diagnoses to special interest groups.

In the context of final determination of instrument suitability by VHA clinicians and managers relative to “fit” with conditions of clinical culture and intended use, the literature indicates that a provisional list of selection criteria would reasonably include:

1. Original purpose congruent with intended VHA use for quality of care tracking and reporting;
2. Multi-dimensional;
3. Acceptable psychometrics;
4. Sensitive to change;
5. Feasible for routine use;
6. Electronic data entry and/or analysis available;
7. Readily interpretable by non-professionals;
8. Low cost.

APPENDIX REFERENCES

Blacker D, Endicott J. Psychometric properties: concepts of reliability and validity. In: Rush AJ, Ed. *Handbook of Psychiatric Measures / Task Force for the Handbook of Psychiatric Measures*. Washington, DC: American Psychiatric Association, 2000.

Eisen SV, Grob MC, Dill DL. Outcome measurement: Tapping the patient's perspective. In: Mirin SM, Gossett JT, Eds. *Psychiatric treatment: Advances in outcome research*. Washington, DC: American Psychiatric Press, 1991: 213-235.

Jenkins R. Towards a system of outcome indicators for mental health care. *British Journal of Psychiatry*, 1990; 157: 500-514.

Kirshner B, Guyatt G. A methodological framework for assessing health indices. *Journal of Chronic Disease*, 1985; 38: 27-36.
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=3972947.

Lohr KN. Outcome measurement: concepts and questions. *Inquiry*, 1988; 25: 37-50.
<http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?db=m&form=6&dopt=r&uid=2966125>.

Lohr KN, Aaronson NK, Alonso J, Burnam MA, Patrick DL, Perrin EB, et al. Evaluating quality-of-life and health status instruments: development of scientific review criteria. *Clinical Therapeutics*, 1996; 18: 979-992. <http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?db=m&form=6&dopt=r&uid=8930436>.

McDowell I, Newell C. Measuring health: a guide to rating scales and questionnaires. New York: Oxford University Press, 1996: 523 Pages.

Newman FL, Hunter RH, Irving D. Simple measures of progress and outcome in the evaluation of mental health services. *Evaluation and Program Planning*, 1987; 10: 209-218.

Owen RR, Thrush CRN. Practitioner and system evaluation. In: Rush AJ, Ed. *Handbook of Psychiatric Measures / Task Force for the Handbook of Psychiatric Measures*. Washington DC: American Psychiatric Association, 2000: 195-220.

Sackett DL, Chambers LW, MacPherson AS, Goldsmith CH, McAuley RG. The development and application of indices of health: general methods and a summary of results. *American Journal of Public Health*, 1977; 67: 423-428. <http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?db=m&form=6&dopt=r&uid=857684>.

Slade M, Thornicroft G, Glover G. The feasibility of routine outcome measures in mental health. *Social Psychiatry and Psychiatric Epidemiology*, 1999; 34: 243-249.

Smith GR. Considerations in choosing, using, and interpreting measures for health care systems. In: Rush AJ, Ed. *Handbook of Psychiatric Measures / Task Force for the Handbook of Psychiatric Measures*. Washington, DC: American Psychiatric Association, 2000.

Switzer GE, Wisniewski SR, Belle SH, Dew MA, Schultz R. Selecting, developing, and evaluating research instruments. *Social Psychiatry and Psychiatric Epidemiology*, 1999; 34: 399-409.
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=10501709.

Wetzler S. The contemporary parameters of assessment in psychopathology. In: Wetzler S, Ed. *Measuring Mental Illness: Psychometric Assessment for Clinicians*. Washington DC: American Psychiatric Press, 1989.



**VA Technology Assessment Program
Office of Patient Care Services (11T)
VA Boston Healthcare System
150 South Huntington Avenue
Boston, MA 02130**

Tel: 617.278.4469 Fax: 617.264.6587
vatap@med.va.gov
<http://www.va.gov/vatap> <http://vaww.va.gov/vatap>

Report released: July 2002