

## THE IAPSRs TOOLKIT: DEVELOPMENT, UTILITY, AND RELATION TO OTHER PERFORMANCE MEASUREMENT SYSTEMS

PAUL ARNS, E. SALLY ROGERS,  
JUDITH COOK, CAROL  
MOWBRAY & MEMBERS OF  
THE IAPSRs RESEARCH  
COMMITTEE

PAUL ARNS, PH.D. IS DIRECTOR OF THE  
AMI/ABLE INTEGRATED SERVICES  
PROGRAM AND AN ASSISTANT  
CLINICAL PROFESSOR OF PSYCHIATRY  
AT HARBOR-UCLA MEDICAL CENTER,  
TORRANCE, CALIFORNIA.

E. SALLY ROGERS, SCD, IS DIRECTOR OF  
RESEARCH, BOSTON UNIVERSITY,  
CENTER FOR PSYCHIATRIC  
REHABILITATION, SARGENT COLLEGE OF  
HEALTH & REHABILITATION SCIENCES,  
BOSTON, MASSACHUSETTS.

JUDITH COOK, PH.D. IS DIRECTOR OF  
THE NATIONAL RESEARCH AND  
TRAINING CENTER ON PSYCHIATRIC  
DISABILITY, CHICAGO, ILLINOIS.

CAROL MOWBRAY, PH.D. IS PROFESSOR  
AND ASSOCIATE DEAN FOR RESEARCH  
AT THE UNIVERSITY OF MICHIGAN,  
SCHOOL OF SOCIAL WORK, ANN ARBOR,  
MICHIGAN.

FOR INFORMATION CONTACT PAUL  
ARNS, PH.D. AT HARBOR-UCLA MEDICAL  
CENTER, 1000 W. CARSON ST., BOX 462,  
TORRANCE, CA 90509.  
TEL: 310/222-1643. FAX: 310/222-5651.  
E-MAIL: [parns@ucla.edu](mailto:parns@ucla.edu)

*The current climate of economic constraint in community-based mental health services has all but required psychosocial rehabilitation (PSR) programs to demonstrate their effectiveness by adopting outcome assessment systems. Agencies of disparate size and varying degrees of sophistication are now responding to mandates to systematically monitor program outcomes. In doing so, however, PSR practitioners and administrators are faced with a bewildering array of competing measures, some of which are impractical, costly, or irrelevant for capturing the outcomes of PSR services (Blankertz & Cook, 1998). To acquaint readers with issues in performance measurement in PSR, this article describes a newly developed outcome measurement system created to document the achievements of service recipients in PSR programs and to monitor progress towards recovery. We discuss principles and logistical issues in performance measurement which are important to PSR agencies and which the Toolkit has been designed to address. Results of the pilot testing of the PSR Toolkit are presented to demonstrate the feasibility of its use, logistical problems in implementation, psychometric properties of the measures, and the Toolkit's sensitivity to change. We discuss next steps in the development of the Toolkit.*

### GENESIS OF THE TOOLKIT

Beginning in the mid-1990s, members of the Research Committee of the International Association of Psychosocial Rehabilitation Services (IAPSRs) took on the challenge of developing a simple, easy to administer, longitudinal assessment tool for PSR programs seeking to compete in managed care environments. The purpose was to have available a tool that would be specific to

psychosocial rehabilitation program goals, and could be used by many diverse PSR agencies to meet the information demands of managed care providers and to improve the performance of their programs. The Research Committee was joined in this effort by the Evaluation Center @ HSRI and the Federal Center for Mental Health Services in the Substance Abuse and

Mental Health Services Administration (SAMHSA), and also utilized extensive expert input and review by leaders in the psychosocial rehabilitation field concerning the selection and definition of key outcomes. The result of these efforts is “The PSR Toolkit.”

A multidimensional outcome protocol, the PSR Toolkit provides a common format and set of definitions for collecting data on an individual’s progress in areas that are critical to rehabilitation goals, including employment, residential status, community tenure, and satisfaction with services. The Toolkit was designed to be readily accessible and easily adoptable by relatively inexperienced PSR agencies who were attempting outcome measurement for the first time. The purpose of this report is to acquaint individuals served by PSR programs, advocates, researchers, and program staff with the Toolkit, its psychometric properties, and its intended utilization. We will describe the Toolkit and the key features of its design, provide background information on other performance measurement approaches as a contrast, present experiences from a multisite pilot test of the Toolkit, and finally discuss lessons learned and next steps in development.

## DESCRIPTION OF THE TOOLKIT

A critical issue in the development of the Toolkit protocol was the desire to have it driven by PSR principles. Thus, the Toolkit has a rehabilitation perspective and a recovery focus. Measured outcomes are those valued by consumers<sup>1</sup> (e.g., work, residential independence) as well as those identified as intended program goals in PSR interventions like experiential learning, role modeling, or skills training. The Toolkit is structured

in modules to permit the measurement of results from individualized programming efforts, as required by the federal legislative mandate for consumer choice in rehabilitation service planning (Cook & Pickett, 1994). The modular nature of the Toolkit means that it can be tailored to an individual’s goals or a program’s orientation. Another characteristic of the Toolkit is that it is functional in its emphasis, assessing characteristics of work behavior, independent living skills, educational attainment, and so forth, rather than assessing psychiatric symptoms for which many valid, reliable measures currently exist. Finally, given the central importance of the service recipient in selecting and controlling his or her own unique rehabilitation service plan and goals, outcomes in the Toolkit are measured at the individual level, rather than at the program level (program level indicators would include, for example, size of waiting lists, number of staff and their qualifications, etc.).

The Toolkit was designed to be congruent with the above principles. In addition, there were a number of logistic and resource considerations which the Toolkit also needed to address. In order to put these considerations and principles into operation, the following guidelines were produced that are used in developing the Toolkit.

### Utilizes a Minimal Dataset

The Toolkit was designed to be usable in a variety of psychosocial programs, with a variety of services, activities, intervention approaches, and resources. In developing the Toolkit, we attempted to determine the minimum data fields necessary to assess the effectiveness of psychosocial programs. This was done by examining existing measures and by obtaining extensive input from PSR leaders in the field, both evaluators and pro-

gram staff. We recognize that for some small programs, even this minimal dataset may seem burdensome. However, these programs could elect to eliminate some of the Toolkit modules. Similarly, some comprehensive programs may wish to add modules to the Toolkit to enrich the number and types of outcomes that can be measured. As a stand-alone instrument, however, the Toolkit is designed to provide critical data across the domains generally of interest to rehabilitation practitioners. Unlike other rehabilitation outcome measures that may have been developed within the context of research studies and where data collection resources were more abundant, the focus of the Toolkit is on programs where the resources for and experiences with data collection were more limited.

### Domain Based

The Toolkit is organized by domains that are of interest to psychosocial rehabilitation practitioners and program administrators. Since most PSR programs concern themselves with the practical day-to-day realities of the people they serve, the focus of the Toolkit is similarly practical. The Toolkit contains *demographic information* that can be used to describe the individuals in the program. In addition, it contains status-based descriptions and measures in these major domains: *employment, educational activity, and residential situation*. It also contains information about *financial status, legal involvement, hospitalization, service satisfaction, perceived quality of life, and the individual’s sense of mastery*. Our goal was to make the Toolkit useful for a wide variety of psychosocial rehabilitation programs including those in which the focus might be on particular domains of role functioning.

1. We use term “consumer” throughout this article while recognizing that others prefer the consumer/survivor, expatient, member or other terms. The use of consumer is meant to be interchangeable with these other terms.

### **Measures Status, Role Functioning, and Rehabilitation Outcomes**

Mental health treatment interventions generally have goals including symptom reduction, service utilization, and relapse prevention. Many performance measurement systems similarly focus on these outcomes. In contrast, the Toolkit was designed to address the goals most often endorsed by rehabilitation practitioners: role functioning and role recovery. Therefore, the measures in each domain are status measures, designed to provide an ordinal level of measurement vis-à-vis individual role functioning in that domain. For example, the employment domain captures whether the individual is engaged in any productive vocational activity, ranging from sheltered work to competitive employment. It also captures other standard indicators of vocational success such as hours worked and wages earned. By examining the individual's change in status, gains in role functioning over time can be assessed with the Toolkit.

### **Based on PSR Values of Recovery, Empowerment, and Mastery**

The Toolkit contains a survey to assess consumers' service satisfaction, empowerment, and feelings of mastery. This section of the Toolkit was added in recognition of the reality that changes in status and role functioning are not the only, or often not even the most important facets of the rehabilitation process. Promoting empowerment and recovery and changes in participants' sense of self may be important precursors to role change, and are important in their own right. Thus, this section of the Toolkit is designed to assess these more subjective indicators of success. These subjective indicators can be combined with the objective status measures in various domains to obtain a more complete picture of participants' recovery.

### **Designed As an Individual Level Instrument**

Many management information and performance measurement systems are designed to collect program-level data to assess program effectiveness, such as the percentage of individuals in a program who are hospitalized in a given period of time. Others are designed to examine process variables or program structure, such as the number of hours of services provided or the ratio of staff-to-clients served. In contrast, the Toolkit was designed as an individual-level data collection instrument, tracking progress and change on a person-by-person basis. However, the data resulting from the individual level of the Toolkit's administration can be "rolled up" or aggregated across an entire program, or can be examined by certain demographic characteristics. For example, a program may wish to examine change over time in vocational status for all individuals living in a certain kind of residence versus those in another type of residence. Or the agency may want to investigate whether its vocational programs are equally effective in helping individuals of different ethnic backgrounds to make vocational gains. The Toolkit would allow such program level interpretations of the data, while still tracking individual level change.

### **Allows Flexible Use to Capture Descriptive Information and to Monitor Change Over Time**

The Toolkit was designed to capture information describing service recipients at baseline as well as to monitor changes over time for those served. Thus, a program can decide what its information needs are and tailor the use of the Toolkit to address those needs. The Toolkit can be administered once, to all participants in the program at a particular point in time to yield a "snapshot" of individuals in the program, both in terms of their demographics and functioning across domains. It can also

be administered repeatedly, and indices of change over time can be computed (e.g., at termination or as part of annual evaluations). The Toolkit provides continuous measures as well as status indicators; in the employment domain, for example, items assess earned income as well as whether or not the individual is competitively employed. Demographic data allow examination of differential effectiveness for subgroups of service recipients. Thus, use of the Toolkit provides multiple sources of information for a program to assess its effectiveness and determine if it has met its performance goals.

### **Minimizes Respondent Burden**

Instruments which require more than 30 minutes administration for the participant or the practitioner probably present an excessive burden. To minimize consumer and PSR staff burden, sections of the Toolkit can be completed by a clerical or administrative worker, depending upon the type and comprehensiveness of records that are routinely kept within the program. Most of the remainder of the Toolkit can be completed by staff. There is only one section that must be completed by the service recipient: the Member Survey which measures service satisfaction, mastery, and quality of life. In pilot testing, we have found that this section is very well received by program participants and can be completed in 5–15 minutes on average. Respondent burden is also minimized because the Toolkit is very portable and requires no special testing equipment.

### **Minimizes Training Time Needed for Administration**

The academic qualifications for administering the Toolkit are minimal because the Toolkit focuses on status measures and factual information. Practitioners can be oriented to the Toolkit and to how the data are captured in a short period of time. This orientation can occur individually or in groups.

### Measures Have High Face Validity

The Toolkit is designed to be domain-based and focused on role status; thus, the majority of information is factual in nature. Since it relies very little on practitioner judgement, the reliability and validity of the assessments are enhanced. In addition, significant effort was expended clarifying terms, such as the operational definitions of employment and residential role functioning. These definitions were evaluated by PSR experts, tested and refined before the pilot testing. Making sure that the definitions are clear and fully understood also increases reliability and validity.

### OTHER PERFORMANCE MEASUREMENT SYSTEMS

In recent years, several other organizations have developed report cards and performance measurement systems which have, with varying degrees of rigor, included indicators related to the delivery and/or impact of mental health services. Some of the more commonly recognized systems today include: (a) the Performance-Based Measures for Managed Behavioral Healthcare Programs (PERMS 2.0, 1998), developed by the American Managed Behavioral Healthcare Association (AMBHA); (b) the National Committee on Quality Assurance's Health Plan Employer Data and Information Set (HEDIS NCQA 3.0, 1997); and (c) the Consumer-Oriented Mental Health Report Card which was developed by the Center for Mental Health Service's (CMHS) Mental Health Statistical Improvement Program (MHSIP, 1997), the American College of Mental Health Administration (1997), and the National Association of State Mental Health Program Directors (1999). Other efforts include one developed by the Foundation for Accountability (FACCT, 2000), the Joint Commission on the Accreditation of

Healthcare Organizations (JCAHO ORYX, 2000), as well as numerous state, local, professional, and corporate initiatives.

While a thorough review of these systems is beyond the scope of this paper, a brief discussion of the PSR Outcomes Toolkit in the context of these initiatives is warranted. It is important to note that the conceptual frameworks, methodology, and goals of these systems vary considerably. As such, no true consensus regarding core behavioral health indicators has developed (Leff & Woocher, 1998), and many of the systems are evolving on an annual basis.

A major difference between the PSR Outcomes Toolkit and many of the systems cited above relates to the intended service recipients under scrutiny. The Toolkit has a clear and explicit emphasis on adults with *severe psychiatric disability* and, more specifically, was developed for use by programs which are providing psychosocial rehabilitation services to such individuals.

Performance measurement systems such as HEDIS and those of JCAHO and FACCT target the broader population of healthcare recipients. With this broadened overall focus, the mental health portions of these systems tend to be somewhat lacking in depth. For example, the HEDIS 3.0 system's mental health component focuses on rates of 30-day ambulatory follow-up to hospitalizations for a variety of psychiatric diagnoses, as well as availability of mental health and substance abuse providers within a system. General behavioral healthcare monitoring systems, such as the PERMS, focus on the broader population of persons receiving mental health and substance abuse services from managed care entities.

Many of the existing performance measurement systems are considerably broader than the PSR Outcomes Toolkit with regard to the range of domains in

which performance is measured. That is, the performance indicators derived from the Toolkit, with a few exceptions, focus on the *outcomes* of services, or the way in which the mental health consumers' life circumstances have changed. While some of the other systems tap outcomes to a greater or lesser extent, most place a much greater emphasis on *process* issues such as access to care, appropriateness or quality in relation to some established standard of service delivery, and service utilization rates; and some have indicators related to the prevention of mental disorders (e.g., MHSIP, FACCT).

Each of the developing performance measurement systems has unique strengths and limitations, as well as particular contexts to which they are best suited. Because of its content and its developmental history, the Toolkit fills a particular niche among the various monitoring systems for assessing indicators most relevant for rehabilitation. However, if these other systems are required by funders or oversight agencies, the Toolkit can be effectively combined with them for a more comprehensive view of the impact of service.

### THE PILOT TEST EXPERIENCE

From 1995 to 1997, the IAPSRs Research Committee conducted a multisite field test of the Toolkit with 13 PSR programs across the country. The primary goals of the pilot study were to examine issues around implementing and using the Toolkit, and to identify and correct problems in the Toolkit definitions and measures. Pilot sites were recruited through a mailing to agencies that were organizational members of IAPSRs. From a pool of approximately 70 agencies that expressed some interest in participation, a sample of 26 agencies was invited to enroll in the pilot study. Given

the primary goals of the study, an attempt was made to select a range of programs that represented a variety of states<sup>2</sup>, sizes, models (residential program vs. “clubhouse”), and localities (rural/suburban/urban). Ultimately, 13 agencies, ranging in size from fewer than 20 to over 600 enrolled consumers, submitted data for the pilot study. Some agencies were self-contained programs, while others were multiservice agencies that had many different programs under their “umbrella.”

The design of the pilot study was to collect baseline Toolkit data using an established start date covering all members currently enrolled in each program. Data included items that required continuous status monitoring (such as hospitalizations over a period of time). The time frame for these continuous status monitoring items was the “previous 6 months.” Sites were to obtain Member Surveys from as many enrolled members as was possible within 2 weeks of the start date. The actual start dates for the different sites varied by several months, depending on local differences in the amount of time needed to obtain necessary approvals, train and engage staff, and so on. Sites collected follow-up data 6 months after their start date, including information on the intervening period where relevant (e.g., hospitalizations). Follow-up data were only collected on individuals who had been enrolled at “baseline” and who remained enrolled in the program at the time of follow-up, to allow examination of individual-level changes over time.

Agencies were provided with Toolkit documentation, scoring instructions, and ongoing technical assistance to pilot-test the instrument. A total of 1384 Toolkits were completed with baseline data on demographics and functioning

**Table 1—Pilot Sample Demographic and Diagnostic Information**

VARIABLE	<i>n</i>		% OF SAMPLE
Gender	1383	Male	58%
Ethnicity	1379	Euro-American	64%
		African-American	28%
		Hispanic	6%
		Asian/Pac. Isl.	1%
		Native American	<1%
		Other	1%
Education	1360	College Graduate +	10%
		Some College	18%
		High School Graduate/GED	44%
		< High School	28%
Marital Status	1377	Never Married	71%
		Divorced/Separated	22%
		Married	4%
		Cohabiting	1%
		Widowed	0%
Primary Diagnosis	1360	Schizophrenia/Other Psychoses	65%
		Bipolar Disorder	9%
		Depression/Other Mood Disorders	15%
		Mental Disorder due to	
		General Medical Condition	2%
		Substance Abuse or Dependence	<1%
		Other Axis I Diagnosis, Personality Disorder or V-Code	7%

*Note:* Because of the way in which the data were reported, it was not always completely clear which diagnosis was considered “Primary” by the sites. When more than one diagnosis was listed, an algorithm was developed which identified the “most likely” primary diagnosis for that individual. More severe, non-substance abuse, Axis I disorders (e.g., Schizophrenia, Bipolar Disorder, Major Depression) were generally assumed to be Primary if present, followed by moderately severe Axis I disorders (e.g., Agoraphobia, Obsessive Compulsive Disorder), then Personality Disorders (Axis II), then Substance Abuse/Dependence, then Adjustment Disorders, and finally “V-Codes”. In the table above, adjustment disorders are combined with “Other Axis I” disorders, but only in the rare instances in which there was no Personality or Substance Disorder present. The software version of the Toolkit removes ambiguities in recording this data.

and at follow-up, a total of 875 Toolkits were completed (63.2%). Two sites did not submit follow-up data due to personnel changes and one due to computer failure. Data were sent periodically to the IAPSRs Research Committee for analysis.

*Sample characteristics.* The mean age of participants was 39.9 ( $SD = 11.5$ ) with a

range of 16–87. Age at first psychiatric hospitalization was 24.5 ( $SD = 9.6$ ). Nearly three-fifth of the individuals were male, 64% white, 28% African American, 5.5% Hispanic, and the remainder other ethnicities. The modal educational level for the Toolkit pilot sample was high school completion, with equal percentages having more than high school and

2. The pilot study was limited to psychosocial rehabilitation agencies in the United States because several Toolkit definitions or categories were not applicable in other countries (e.g., income categories in the financial domain).

less than high school educations (28%). Nearly three-quarters of the sample had never married and nearly one-quarter were separated or divorced. The majority had a diagnosis of schizophrenia; about a quarter were diagnosed with a unipolar or bipolar affective disorder. See Table 1 for more detailed information.

## RESULTS OF THE PILOT TESTING

Pilot test results are provided here to give the reader information on the psychometric characteristics of the Toolkit data, some idea of the feasibility of the data collection effort, and examples of some of the indicators that can be derived from the data. This presentation is not intended to be an exhaustive or comprehensive description of the findings nor of the complexities of the data analysis. We refer the reader to the Toolkit Pilot Study Final Report, available from IAPSRS,<sup>3</sup> for a more complete review.

*Participant acceptance of the data collection.* The design of the pilot test called for service recipients to complete member surveys within 2 weeks of the staff completing the Toolkit demographics and functioning domains. Consumer participation was high in most agencies, with 82% of the individuals requested to complete baseline surveys doing so.

*Distribution of assessment data.* On most of the measures, this diverse sample from multiple PSR agencies showed an adequate distribution of scores. An adequate distribution of and variation in scores are important since such findings indicate that differences among subgroups are more likely to be detected, and that a range of more sensitive statistical tests may be used. For example, the

proportion of individuals having any hospitalization in the 6 months before the baseline assessment ranged from 0% to 30% across programs. The mean number of days of hospitalization per episode was 5.04 ( $SD = 19.94$ ). The sample also showed variation in questions about residential status: 26% of the sample lived independently, 29% were in supported housing, 28% in supervised facilities, and 12% lived in a supervised nonfacility. In terms of employment status, many of the employment options showed a full range across study sites with 0–100% not working, 0–80% in unpaid work, and 0–24% in independent competitive employment. For monthly income, 36% of the sample received less than \$500; 52% received \$500–\$999; and 11% received over \$1000 per month.

Some of the status options did have a skewed distribution and limited range. For the most part, though, these options represented states that are expected to be low for those receiving PSR services (i.e., living in prison, 0–1%; being homeless, 0–9%; living in an institution, 0–5%; having been arrested, 3.9% overall average with a range of 0–30%). One measure of concern may be the percent of individuals with any substance abuse treatment episodes, 0–1.4%, that appears to be extremely low. The Toolkit assessment of substance abuse *treatment* episodes is probably not an adequate measure of the magnitude of substance abuse *problems* in this population.

*Sensitivity to change.* The Toolkit was useful for delineating change between assessments in several domains. For example, data collected indicated that 16.5% of the sample had had a psychiatric hospitalization between episodes. Examining data for those completing both baseline and follow-up assess-

ments ( $N = 871$ ), we found a significant decrease in the percentage of the sample with any psychiatric hospitalizations (from 17.5% to 13.0%; chi-square = 7.21,  $p < .007$ ). We were also able to calculate changes in residential status using information from the 483 individuals on whom residential data was submitted at both assessments (one large site excluded this domain from their second assessment). The only significant change in residential status over the 6 month period was a decrease in homelessness (from 2.3% to .6% of the sample, chi-square 4.64,  $p < .03$ ). The fact that other residential categories did not show a significant change over this relatively short period may well be a positive indication of residential stability.

In examining change over time in employment status (limited to  $n = 842$  with both baseline and follow-up employment information), Toolkit data showed that there was an increase in individuals in assisted or independent competitive employment (from 12% at baseline to 17% at follow-up; chi-square = 8.2,  $p < .004$ ), and a decrease from baseline to follow-up of those in any kind of TEP or community-based employment with on-site support from a rehabilitation professional (13% to 8%; chi-square = 9.95,  $p < .002$ ). The percent of the sample without work remained essentially the same at 64%. Overall, we were satisfied with the sensitivity of these status measures to capture changes over time.

*Member Survey.* The Member Survey originally consisted of four subscales (Subjective Quality of Life, Mastery, Empowerment, and Program Satisfaction), each containing five items, answered on a four-point scale. The items for the Member Survey were developed or were derived from existing measures of similar constructs, selected

3. The IAPSRS Toolkit and the Toolkit Pilot Study Final Report can be obtained by contacting the IAPSRS headquarters at (410) 730-7190.

for their appropriateness to PSR principles. In the pilot study, the scales were to be administered at the time of baseline data collection and again 6 months later. In addition, a small sample from three sites was readministered the survey approximately a week after the original administration to provide an assessment of “test-retest” reliability. Retest reliability is a measure of how stable the measure is over a short period of time. In other words, do individuals provide about the same responses on the same questions over a short period of time (in this case, between 3 and 10 days after the first administration)?

A factor analysis—using principal component factor analysis with promax rotation—revealed that the survey could be best conceived as consisting of three subscales: Quality of Life, Mastery, and Program Satisfaction. Results from these three subscales are reported. The test-retest reliability correlations ( $N = 12$ ) were: .85 for Quality of Life, .57 for Mastery, and .82 for Program Satisfaction. Coefficient alphas were calculated to determine the internal consistency of the scales. At baseline, the coefficients were: .75 for Quality of Life, .65 for Mastery, and .81 for Program Satisfaction ( $N$  at least 1023). The data indicate that the Mastery Subscale did not achieve satisfactory levels of test-retest reliability or internal consistency; however, all other subscales were satisfactory.

*Anecdotal findings from the pilot testing.* Numerous suggestions were made by pilot sites to clarify coding decisions for hospitalization, residential and employment status, as well as to capture program status (for example, whether the person is an “active” or “semi-active” member of the agency). Given the proliferation of residential models and alternatives to hospitalization across the sites, it was clear that definitions of treatment and service models in the Toolkit had to be expanded and refined.

A similar problem occurred with coding of employment data where individuals who work intermittently or casually could not easily be coded. Some pilot sites also emphasized the importance of being able to continuously track employment status, as opposed to the “snapshot” approach utilized in the Toolkit pilot version, such as collecting employment data at one point in time and then another 6 months later.

A limitation of the pilot was that the “baseline” data, that was collected at the pilot study “start date,” rarely represented a true baseline in the sense of being a member’s status at the time of enrollment in the program. Program participants had been enrolled in their respective programs for varying lengths of time, ranging from a few weeks to several years. This probably decreased the degree of change that was observed among the members of these PSR programs over time, relative to what might be expected in a true prepost examination of program impact.

The pilot study identified problems in obtaining approvals to initiate the data collection and in implementing the data collection effort along with effective remedies. These included the need for a well developed implementation plan, the importance of obtaining support for the Toolkit from line staff and administrators, ways to negotiate clearances related to individuals participating in research protocols, as well as the importance of careful data monitoring and follow-through. The pilot test pointed out the need for some PSR agencies to have technical support available to them when first considering and using this outcome measurement system. From the pilot testing we also became aware of the amount of effort required to obtain complete and accurate information. This was particularly true of sites that had not engaged in data collection efforts in the past. Even though the PSR

Toolkit was designed to be minimally burdensome on practitioners, the pilot study results showed the importance of strategizing about the use of the Toolkit, and how to incorporate line staff into decision-making and interpretation of the data in order to increase their investment in the process. Some programs have found “participatory outcome teams” useful to monitor the implementation of the Toolkit and to problem-solve around difficulties and barriers encountered.

### NEXT STEPS WITH THE TOOLKIT

To address the implementation issues which surfaced in the pilot study, a number of steps have been taken. First, problematic definitions were revised. Second, to further minimize burden, a software version of the Toolkit has been developed and is available. Further, the software is accompanied by access to technical assistance on implementation and assessment issues. Third, workshops and training conducted at IAPSRs annual conferences provide more information to interested agencies on implementation issues and on successful techniques agencies have used to resolve them. This includes suggestions as to how to begin data collection and how analyses and output can deal with baseline data collected on consumers who have been enrolled for some time.

### PERFORMANCE INDICATORS IN THE PSR OUTCOMES TOOLKIT

The Toolkit itself simply provides a common format and definitions for the collection of data across several relevant outcome domains. In order to serve as a performance measurement system, performance indicators need to be derived

**Table 2—Possible Performance Indicators Derived from the PSR Toolkit<sup>1</sup>**

**Hospitalization Domain**

- % with no psychiatric hospitalizations during year
- % with > 2 psychiatric hospitalizations during year
- % with any medical hospitalizations during year
- % psychiatric hospitalization episodes which are “voluntary”
- % rehospitalized within 30 days of a hospital discharge
- % psychiatric hospital stays lasting > 10 days
- % with over 30 psychiatric inpatient days over year
- Median length of stay for hospital admissions

**Residential Domain**

- % living independently (level 7 in “residential status definitions”)
- % living in independent or supported housing (level 6 or 7)
- % showing increase in residential status time 1 to time 2
- % in “stable” housing (no changes over year) \*
- % in institutional settings (levels 1 or 3)
- % currently homeless
- % with any homeless episodes during year \*

**Employment Domain**

- % in paid employment (status 3 – 11) currently
- % in any paid employment during year \*
- % in any work-type activity (status 2-11) currently
- % in “competitive” job settings (status 10-11) currently
- % in competitive employment during year \*
- % of those working who received employer paid health insurance benefits
- % working > 20 hours per week in paid job
- % with weekly employment earnings > \$150
- % with > 60 days paid employment during year \*
- Mean and Median job tenure among those who worked during year \*

**Education Domain**

- % currently enrolled in education program in community
- % in educational setting > = 8 hours per week
- % enrolled in education program during past year \*
- % in “productive” community activity (school or employment 2-11)

**Financial Domain**

- % with total monthly income > \$700
- median monthly income
- % receiving SSI or SSDI
- % receiving monetary support from family
- % with any employment earnings

1 With the exception of the Hospitalization and Legal domains, the “paper and pencil” version of the PSR Toolkit captures data cross-sectionally. The software version allows status changes and events to be captured across all domains over time. Indicators marked with an asterisk (\*) are possible with the software version.

*(table continued on next page)*

from the data that are collected. Based on the Toolkit pilot study, a list of several possible performance indicators for each domain has been developed (see Table 2).

When selecting and utilizing performance indicators, several caveats are warranted. First, programs must be attentive to “case mix”<sup>4</sup>. That is, the specific characteristics of the individuals served by a psychosocial rehabilitation program may vary considerably with regard to diagnoses and comorbid conditions, symptom severity, community living skills, and so forth. Thus, between program comparisons should be attempted cautiously and need to be sensitive to “case mix”—that is, assuring that outcome comparisons involve similar groups of service recipients or, when dissimilarities do exist, controlling for these through matching or statistical procedures. Additionally, the characteristics of individuals served by a particular program may influence reasonable expectations about “adequate” performance on certain outcome measures. For example, a group of consumers with very little vocational history or advanced education would not be expected to change as much in increased earnings as individuals with some college education and substantial work histories. “Case mix” can also shift over time *within* a program; possible changes in the aggregate characteristics of service recipients should be examined and, if noted, accounted for in analyses of agency performance over time.

A related concern is the importance of examining the representativeness of the individuals on whom data are collected when examining a performance indicator. If a performance indicator (e.g., mean program satisfaction score) is derived from a subset of individuals in a program, one must be careful that no

4. While we prefer not to refer to the recipients of psychosocial rehabilitation services as “cases”, we have used the term “case mix” in this article due to its specific meaning in the evaluation and research literature.

systematic bias exists that could influence the generalizability of findings (e.g., most of the people who appear to be dissatisfied with the program refused to complete a survey).

Another important caveat is the need to be attentive to the particular goals and objectives of a psychosocial rehabilitation program. For example, a program with a strong vocational component may place relatively greater emphasis on employment and educational outcome indicators than an agency that primarily focused on providing stable housing. Thus, attempts should be made to focus on outcome indicators that match program goals.

Programs exist within a broader social and political system. Thus, the environmental context can also influence outcomes in various domains as much as the efforts of the program and the nature of the people being served.

Hospital closures, new legislation, and changing admission policies can have a direct impact on hospitalization related outcome indicators. Similarly, availability of housing options in a community, and local economic conditions can influence consumer outcomes and therefore agency performance in residential and in employment domains. These are all relevant issues that agencies should keep in mind when comparing data across programs or within a program over time.

The set of proposed performance indicators shown in Table 2 should be considered representative of the possibilities that agencies may select as outcomes, rather than comprehensive. In using these indicators, evaluators should consider the appropriateness of various time frames and categories. Most of the indicators are in the form of "head counts" (e.g., percentage of individuals who were hospitalized over

**Table 2—Possible Performance Indicators Derived from the PSR Toolkit' (continued)**

**Legal Domain**

- % with any arrests during year
- % with > 1 arrest during year
- % who spent any time in jail during year
- % with > 30 jail days during year
- % currently on probation or parole
- % who were victims of a violent crime during year
- % who were victims of a nonviolent crime during year

**Subjective Indicators from the Member Survey**

- Mean Program Satisfaction Score (items 3, 4, 5, 7, 10, 14, 15, 17, 20)<sup>2</sup>
- Individual item analysis: % who "agree" or "strongly agree"
  - Social Support – (item 6) "Overall, I have a good relationship with members of my family"; (item 13) "I have an active social life"
  - Residential Satisfaction – (item 11) "I am happy with my current living situation"
  - Physical Health Concerns – (item 19) "I am in good physical health"

<sup>2</sup> Pilot study results indicated that a fairly robust Program Satisfaction scale can be derived from the noted items. Other subscales originally envisioned were less psychometrically adequate. Potentially useful information can still be obtained on particular topics by examining other survey items individually.

some period of time), primarily because measures such as the mean number of hospital days over the time period can be substantially skewed by a few "outlying" values. The Toolkit Pilot Study Final Report contains a more detailed description of the important issues in using these indicators.

## THE TOOLKIT AND PSR PROGRAM ACCOUNTABILITY

The advent of managed care means that many PSR programs must demonstrate their effectiveness relative to other kinds of services for individuals with severe psychiatric disability<sup>5</sup> (Anthony, 1992). The Toolkit's focus on rehabilitation outcomes allows it to provide the kinds of information required by national professional accrediting bodies such as the Commission on the Accreditation of Rehabilitation Facilities (CARF). In addition, state mental health authorities may

require that the agencies they fund collect information on consumer satisfaction with services and program-wide rates of key indicators such as rehospitalization and employment (Speer, 1996). Thus, use of the Toolkit can serve multiple purposes, helping to meet the many requests PSR agencies now find themselves facing from accreditation, funding, and utilization review bodies.

## CONCLUSIONS

Psychosocial rehabilitation programs are under increasing pressure to demonstrate that they improve the lives of people with psychiatric disabilities. Funding bodies are stepping up their demands for accountability, and accreditation bodies increasingly require program evaluation data. Frequently PSR programs are at a loss about how to develop a program evaluation initiative that can serve these needs.

5. We use the term severe psychiatric disability interchangeably with severe and persistent mental illness or severe mental illness.

The IAPSRs Toolkit is based on rehabilitation and recovery concepts and values, and allows for the derivation of numerous performance indicators across several outcome domains as described in this article. The Toolkit was developed to help PSR programs meet the demand for accountability and program evaluation data. It was designed to be a brief, but structured instrument. Piloting experience suggests that many programs can implement the Toolkit with relatively minimal burden on staff, consumers, or program resources. The Toolkit can be paired with other performance measurement systems presently being used by hospitals and outpatient clinics, such as HEDIS and PERMS. These performance measurement systems tend to focus on process measures or clinical management rather than on rehabilitation outcomes, and thus often cannot alone serve the needs of PSR programs.

The Toolkit has been extensively pilot tested and revised. Indications are that most measures on the Toolkit have sufficient validity and reliability, and are appropriately sensitive to change. The Toolkit has been in use for over a year in the province of Ontario, Canada, and is being used by numerous PSR programs in the United States. It has been developed into an easy to use software application available from IAPSRs<sup>3</sup>. The pilot testing revealed several problems with the Toolkit, including the performance of the Member Survey, definitional and coding difficulties, and problems related to implementing a program-wide data collection effort. These problems have been addressed in the current Toolkit version. The next steps for the PSR Toolkit include developing additional outcome modules and refining the Member Survey so that the Toolkit can be used to enhance the breadth, reliability, and comprehensiveness of outcomes that are captured.

The IAPSRs Research Committee welcomes input from agencies using the Toolkit. We also welcome inquiries from agencies interested in beginning implementation of the Toolkit. We see this as an evolving effort to meet the internal management needs and quality improvement goals of PSR agencies, along with increasing performance and accountability demands of funders and oversight agencies.

## REFERENCES

- American College of Mental Health Administration (1997). *The Sante Fe summit on behavioral health*. Pittsburgh, PA: Author.
- American Managed Behavioral Healthcare Association (1998). *Performance measures for managed behavioral healthcare programs* (PERMS 2.0). Washington, DC: Author.
- Anthony, W. A. (1992). Managed mental health care: Will it be rationed care or rational care. *Psychosocial Rehabilitation Journal*, 16, 120–123.
- Blankertz, L., & Cook, J. A. (1998). Choosing and using outcome measures. *Psychiatric Rehabilitation Journal*, 22(2), 167–174.
- Cook, J. A., & Pickett, S. A. (1994). Recent trends in vocational rehabilitation for people with psychiatric disability. *American Rehabilitation*, 20(4), 2–12.
- Foundation for Accountability (FACCT). *Facct quality measures*. Portland, OR: Author.
- IAPSRs Research Committee (1998). *Measuring psychosocial rehabilitation outcomes: Toolkit pilot study final report*. Columbia, MD: International Association of Rehabilitation Services.
- Joint Council on the Accreditation of Health Care Organizations (2000). *1999–2000 comprehensive accreditation manual for behavioral health care*. Oakbrook, IL: Author.
- Leff, H. S., & Woocher, L. S. (1998). Performance measurement systems: A conceptual approach. *Outlook*, 1998 (Summer) 1-3. Cambridge, MA: *The Evaluation Center @ HSRI*.
- Mental Health Statistical Improvement Program (1997). *Consumer-oriented mental health report card*. Washington, DC: Center for Mental Health Services, Substance Abuse, and Mental Health Services Administration.
- National Association of State Mental Health Programs (1999). *NRI behavioral health care performance measurement system*. Lexington, Kentucky: Author.
- National Committee on Quality Assurance (1997). *Health plan employer data and information set (HEDIS 3.0)*. Washington, DC: Author.
- Speer, D. C. (1996). Mental health services outcome evaluation. *Clinical Psychology: Science and Practice*, 3(2), 105–129.