the EVALUATION@HSRI CENTER

Psychiatric Rehabilitation Fidelity Toolkit

November 2000

Prepared by:

Gary Bond, Jane Williams, Lisa Evans, Michelle Salyers, Hea-Won Kim and Heather Sharpe
Department of Psychology
Indiana University-Purdue University
Indianapolis, IN

H. Stephen Leff Ph.D.
the Evaluation Center@HSRI
Cambridge, MA

Human Services Research Institute
2269 Massachusetts Avenue
Cambridge, MA 02140
www.tecathsri.org

U.S. Department of Health and Human Services
Substance Abuse and Mental Health Services Administration
Center for Mental Health Services
www.samhsa.gov

the EVALUATION@HSRI CENTER

This Toolkit is one of a series of such kits commissioned by the Evaluation Center@HSRI. The Center is supported by a cooperative agreement with the Center for Mental Health Services, Substance Abuse and Mental Health Services Administration. The mission of the Evaluation Center is to provide technical assistance related to the evaluation of adult mental health systems change.

The Center offers seven programs all of which are designed to enhance evaluation capacity. ***The programs are:*** **the Consultation Program**, which provides consultation tailored to the needs of individual projects; **the e-Community Program**, which provide a forum for ongoing dialogue via electronic conferencing; **the Toolkits & Materials Program**, which provides evaluators with tested methodologies, instruments and original papers on selected topics and identifies relevant literature in the field; **the e-Learning Program**, which supplies online courses and in-person training; **the Multicultural Program** that provides technical assistance with respect to evaluation of mental health services and systems for racially, ethnically and culturally diverse persons; **the Conferences Program** designed to inform our audience of events in which issues related to evaluation research are discussed; and the **Evidence-based Practices Program,** which assists in identifying evidence-based practices and moving promising interventions to evidence-based service.

The Toolkits are designed to provide evaluators with complete descriptions of methodologies and instruments for use in evaluating specific topics. Based on information from a needs assessment study conducted by the Center and on feedback from evaluators in the field, we have identified a number of important topics that evaluators are frequently interested in examining. Expert consultants have been engaged to review the background of these topics and to compile Toolkits that provide evaluators with state-of-the-art evaluation techniques to use in their own work.

The Evaluation Center@HSRI has also established an online Forum for discussing issues surrounding its Toolkits as well as other issues related to mental health service evaluation. This forum will provide an electronic venue for Toolkit users to share their expertise and experiences with the Toolkits. If you would like to participate in a user group, please visit and e-forum area of our website, www.tecathsri.org.

We hope that this Psychiatric Rehabilitation Fidelity Toolkit will be helpful to those evaluators who are interested in measuring fidelity..

H. Stephen Leff, Ph.D.
Director

Virginia Mulkern, Ph.D.
Associate Director

Acknowledgement

All the authors are affiliated with the Department of Psychology, Indiana University-Purdue University Indianapolis or Human Services Research Institute (HSRI). Work on this paper was supported by a grant from the HSRI and Grant MH00842 from the National Institute of Mental Health (NIMH). We thank Dawna Phillips from HSRI for her feedback and support in completing this project. *Correspondence concerning this toolkit may be addressed to: Gary R. Bond, Department of Psychology, Indiana University-Purdue University Indianapolis, 402 North Blackford Street, Indianapolis, IN 46202-3275; Phone: (317) 274.6752; Fax: (317) 274.6756; e-mail: gbond@iupui.edu.*

Contents

Preface

When we undertook this project at the invitation of Steve Leff of the Evaluation Center@HSRI, we optimistically assumed that we could produce a document representing state-of-the-art knowledge and a helpful guide for researchers, evaluators, and administrators. As we completed our project, however, we became painfully aware that that our task was overly ambitious, transecting issues in conceptualization of program models, measurement theory, and evaluation of program models. In none of these areas is the knowledge base static. In particular, new studies are being published on fidelity measurement every month. In the course of our work, we discovered that numerous researchers around the U.S. and abroad are developing fidelity measures, using a variety of strategies. Our conversations with researchers have revealed a lack of norms about what fidelity is and how fidelity should be measured. This ambiguity indicates a need for a coherent statement and represents an important opportunity.

We struggled with defining the audience to which this toolkit is aimed. It is intended for researchers and program evaluators who understand basic concepts of measurement. Portions of the toolkit are intended to reach a general audience, including program administrators and others who seek to monitor psychiatric rehabilitation programs. This broader audience may be especially interested in the Appendix, which provides a compendium of instruments in use and under development.

At present, models of psychiatric rehabilitation are incomplete, our body of evidence-based knowledge is meager, and methods of measuring fidelity are crude. With emerging developments, we hope and expect that this document will rapidly become obsolete. We ask for your feedback and hope that you use whatever parts of this document that may be of value for your particular endeavors. Also, we assume that there are other fidelity scales in use of which we are unaware and would like to hear about any fidelity projects under way that we have not included here.

Enclosed you will find a card to sign up for a user group for this toolkit. The authors of the toolkit and the Evaluation Center@HSRI hope there will be sufficient interest to form such a group. Sharing information among group members would be one way to support further developments in measuring fidelity. With enough interest, a listserve may also be established for persons interested in measuring fidelity. If you wish to be informed about these and other activities in the area of measuring fidelity, please return the enclosed card or contact the Evaluation Center@HSRI, 2336 Massachusetts Avenue, Cambridge, MA 02140; Phone: (617) 876-0426; Fax: (617) 497-1762; e-mail: @hsri.org. Web Site: http://tecathsri.org.

Overview

Purpose and Scope of the Toolkit

The primary purpose of this toolkit is to present a working guide for the development of fidelity measures to be used in assessing the implementation of psychiatric rehabilitation program models. Chapter 1 describes the origins of fidelity measures and discusses their research and practical applications. Chapter 2 reviews current models in psychiatric rehabilitation. Chapters 3-5 provide a detailed guide for developing fidelity measures. The Appendix gives examples of instruments currently in use for psychiatric rehabilitation.

Fidelity refers to the degree to which a particular program follows a program model. In turn, a program model refers to a well-defined set of prescribed interventions and procedures. Program models specify such things as the types and amounts of services persons should receive, the manner in which services should be provided, and the administrative arrangements necessary to support service delivery. Fidelity measures are tools to assess the adequacy of implementation of program models. In essence, fidelity measures quantify the degree to which the elements in a program model have been adequately implemented.

Increasingly, measures of program model fidelity have become standard requirements in mental health services research (Bond, Evans, Salyers, Williams, & Kim, 2000; Heflinger, 1996; Henggeler, Pickrel, & Brondino, 1999). Despite this attention, no systematic body of theory and research on fidelity has yet appeared in the mental health services area. This toolkit is intended to begin to address this gap.

Psychiatric rehabilitation refers to services and programs intended to help adults with severe mental illness attain optimal integration into normal adult roles in the community. As described in Chapter 2, psychiatric rehabilitation approaches typically are classified according to areas of role functioning, namely, community integration, independent living, employment, academic achievement, social relationships, communication skills, and family relationships (Dincin, 1995). Corresponding to this list are the following broad categories of psychiatric rehabilitation approaches: case management (community integration), residential programs (independent living), vocational programs (employment), supported education (academic achievement), drop-in centers (social relationships), skills training (communication skills), and family psychoeducation (family relationships). Within each of these categories are specific program models, which refer to particular approaches to helping individuals achieve optimal functioning in one or more domains. Models of case management, for example, include assertive community treatment, the strengths model, the rehabilitation model, and the brokered model (Solomon, 1992).

Contents of the Toolkit

Chapter 1. History and Uses of Fidelity Measures

We give a rationale for the development and use of fidelity measures, beginning with a review of the origins of fidelity measurement within the psychotherapy and psychiatric rehabilitation literatures. We then offer examples of the ways that fidelity measures have been and could be used, from both scientific and practical perspectives.

Chapter 2. Mapping out the Domains of Psychiatric Rehabilitation

We review the domain of psychiatric rehabilitation to provide context for issues in fidelity measurement. The specific types of programs we discuss include case management, vocational programs, supported education, residential programs, skills training, drop-in centers, and family psychoeducation.

Chapter 3. Preparing for Scale Development

In this chapter we give a broad overview of the steps to develop a fidelity scale and discuss more in depth the decision making process of the first steps of developing a scale. These first steps include defining the purpose of the fidelity scale, identifying similar scales in use, and identifying the dimensions of program models.

Chapter 4. Scale Development

This chapter takes the reader through the step by step process of developing a fidelity scale including choosing data sources, developing items, and planning the data collection process.

Chapter 5. Piloting the Scale

We discuss the importance of piloting the fidelity scale to establish the reliability and validity of the measure. The process of piloting, assessing psychometric properties, and revising the scale is described.

Chapter 6. Conclusions and Recommendations

In this chapter we review the importance of and need for continued development and refinement of fidelity measures. We also discuss the possible challenges and obstacles associated with developing fidelity measures.

Appendices. Fidelity Instruments

We present a list of fidelity measures developed for use in psychiatric rehabilitation and we describe more extensively those that have documented information on utility, reliability, validity, and other available information. The scales presented in the Appendix are at various stages of development and use. We have attempted to include examples that provide a sampling of the many uses and developments of fidelity scales. We include scales based on widely-used program models and scales based on specific psychiatric rehabilitation programs.

Chapter 1. History and uses of Fidelity Measures

Origins of Fidelity Scales

The origins of fidelity measurement can be traced to the psychotherapy literature. The importance of defining and measuring elements of psychosocial intervention approaches first began to be recognized in the 1960s, when psychotherapy researchers discovered the impossibility of sorting out the methodological and interpretative problems in early outcome studies (Moncher & Prinz, 1991; Waltz, Addis, Koerner, & Jacobson, 1993). Early psychotherapy research assumed that different psychotherapy approaches were fundamentally different and that therapists from the different schools (e.g., client-centered, psychodynamic) conducted their therapy sessions in a distinctive manner. The evidence mounted that these assumptions were not true (Eysenck, 1952). Psychotherapeutic approaches were poorly defined, with great variation among practitioners. The goal of building a cumulative body of knowledge in science was impossible without defining the interventions more rigorously.

Client-centered therapy was among the first schools of psychotherapy to systematically examine its therapeutic methods and techniques (Rogers, 1951; Rogers, 1957). These constructs were used as the basis for the development of process rating scales for client-centered therapy – what we would call fidelity measures (Bond et al., 2000).

Fidelity measurement accelerated the maturation of psychotherapy research by making standardized treatments possible and by providing methods to document differences between different forms of treatment. The measurement of fidelity developed in two directions associated with two related methodological issues (Moncher & Prinz, 1991). The first, referred to as treatment integrity, concerns the degree to which a treatment condition is implemented as intended. The second, referred to as treatment differentiation, refers to "whether treatment conditions differ from one another in the intended manner such that the manipulation of the independent variable occurred as planned" (Moncher & Prinz, 1991).

In a similar vein, Waltz et al. (1993) suggested four types of therapist behaviors to be considered in assessing adherence: (a) Behaviors that are unique and essential to the model; (b) Behaviors that are essential but not unique; (c) Behaviors that are compatible with the model but are neither essential nor unique; and (d) Behaviors that are prohibited. To apply this distinction to a psychiatric rehabilitation program model, one might develop the following responses for assertive community treatment (ACT) as compared to a traditional case management program: (a) shared caseloads; (b) assertive outreach; (c) recreational activities, and (d) exclusive office-based interventions. However, as described below, there is not yet consensus on each of the four behaviors for mental health program models, even one so well described as ACT.

In addition to the development of fidelity measures, the notion of operationally defining program models had another set of implications for research and practice, namely, that effective treatments could be disseminated widely. Elements of an effective model could be systematically described and implemented by training and supervising staff in the application of the model principles and techniques. Following the client-centered therapy example, the standard in psychotherapy has been to develop treatment manuals (also called practice manuals), which provide detailed descriptions of how treatment services should be organized and how providers should perform their responsibilities.

Psychotherapy researchers soon realized that as important as fidelity is for describing program implementation, other facets of implementation not subsumed under fidelity are also critical (Calsyn, 2000). Certainly, practitioner competencies are essential for the success of any intervention (Waltz et al., 1993). Without competent staff to implement a model, standards of program implementation have little practical meaning. Many a program has fallen short of the ideal because of inexperienced staff. Another such element has been referred to as "the dose-response function" (Lipsey, 1990) or the related concept of the "strength" of the intervention (Scott & Sechrest, 1989). The concept of a dose-response is borrowed from drug research and relates to how much a person responds in relation to the amount of intervention received. In a series of elegant meta-analyses, Howard, Kopta, Krause, and Orlinsky (1986) showed that the number of psychotherapy sessions was exponentially related to symptom relief, with the first few sessions having substantial impact, but later sessions asymptoting in their incremental benefit. An application of the dose-response function in psychiatric rehabilitation would be the amount of improvement seen as a function of the number of contacts in an intensive case management program. A fidelity measure might stipulate a specific minimum level of contact, whereas a dose measure would quantify levels of intensity even beyond the minimum acceptable level.

Fidelity in Psychiatric Rehabilitation

Like the early psychotherapy literature, the psychiatric rehabilitation literature generally has lacked even basic descriptions of program models (Bond et al., 2000). In a review of the community support program (i.e., psychiatric rehabilitation) literature, Brekke (1988) found that only one of 33 studies satisfied his criteria for a complete program description. Case management reviews also have found ambiguities in program model descriptions (Gorey et al., 1998; Latimer, 1999b; Marshall & Creed, 2000; Marshall, Lockwood, Green, & Gray, 1998; Mueser, Bond, Drake, & Resnick, 1998), as have reviews of vocational rehabilitation for people with SMI (Bond, Drake, Becker, & Mueser, 1999a). In the psychiatric rehabilitation field, program labels abound and often are used with different meanings. To take one such example, intensive case management, assertive case management, mobile treatment teams, continuous treatment teams, and assertive community treatment, are sometimes intended to mean the same thing, sometimes something different (Marshall & Creed, 2000). Although experts may agree to some extent what is meant by each of these program labels (McGrew & Bond, 1995; Schaedle & Epstein, 2000), there is no unanimity on elements for these models that are unique, essential, compatible, and prohibited.

Measurement of fidelity in psychiatric rehabilitation has not been completely ignored, however. In a pioneering effort, Paul and his colleagues (1977) developed a hospital-based social learning approach with an intricate method for assessing program fidelity. In another early effort to measure fidelity, Anthony, Cohen, and Farkas (1982) identified 10 "essential ingredients" for determining if a program followed psychiatric rehabilitation principles. They later developed a formal coding system for assessing partial hospitalization programs on these principles (Fishbein, 1988). However, neither of these measurement approaches was widely adopted.

Stein and Test's (1980) landmark study on the ACT model was important not only for demonstrating an effective alternative to hospitalization, but also for the clarity of its program model. In a paper that anticipated trends in the health care field, Test and Stein (1976) outlined a set of practice guidelines essential for implementing their model. Their work laid the foundation for a process study of ACT (Brekke & Test, 1987) and the later development of ACT fidelity scales (e.g., McGrew, Bond, Dietzen, & Salyers, 1994; Teague, Bond, & Drake, 1998; Teague, Drake, & Ackerson, 1995).

Using a diversity of data sources, Brekke showed that systematic measurement of theoretically relevant dimensions of psychiatric program models was possible (Brekke, 1987; Brekke & Aisley, 1990; Brekke & Test, 1992; Brekke & Wolkon, 1988). His work highlighted the feasibility of model differentiation through empirical methods. Brekke and Test (1992) empirically documented differences in program practices in an ACT program, a psychosocial rehabilitation clubhouse, and a hybrid program based loosely on a Fairweather lodge.

Progress in developing psychiatric rehabilitation fidelity measures has been hampered by several factors. One major factor has been the lack of well-defined models. With the exception of ACT (Allness & Knoedler, 1998; Stein & Santos, 1998), skills training (Wallace, Liberman, MacKain, Blackwell, & Eckman, 1992), the Individual Placement and Support (IPS) model of supported employment (Becker & Drake, 1993), and some family intervention approaches (Mueser & Glynn, 1999), treatment manuals have been rare in psychiatric rehabilitation. The development of fidelity scales is much easier with detailed manuals.

The complexity of psychiatric rehabilitation services poses a great challenge to fidelity measurement. Whereas in psychotherapy, the focus is on therapist behaviors, model fidelity in psychiatric rehabilitation typically concerns not only practitioner behavior but also structural aspects of a program (e.g., caseload size, staff qualifications), location of services (e.g., in community settings), and "behind the scenes" activities (e.g., integration of treatment and rehabilitation). Manualizing counseling and psychotherapy often involves minute-to-minute specification of specific therapist interventions, whereas practice manuals for psychiatric rehabilitation models inevitably must be conceptualized at a more macro level. Psychiatric rehabilitation models are inherently difficult to manualize, because the interventions occur in multiple settings, with multiple providers and recipients, and involve diverse activities that go far beyond a counseling setting.

Other Trends Related to Fidelity in Psychiatric Rehabilitation

In addition to fidelity measures, other tools have been increasingly used to improve treatment integrity and differentiation in psychiatric rehabilitation. These include practice guidelines, program certification, report cards, and general measures of work environment. We describe each of these trends and their relationship to fidelity measurement.

Practice guidelines

Along with fidelity measures, practice guidelines have begun to receive attention in the psychiatric rehabilitation literature. Like treatment manuals, practice guidelines explain what services to provide, to whom, and how. Unlike treatment manuals, practice guidelines usually are not "model specific," but rather indicate recommended services targeted to people with a specific illness (e.g., schizophrenia).

Practice guidelines began to receive serious attention in medicine in the 1980s (Eddy, 1990a). Guidelines specific to schizophrenia and to severe mental illness have been developed by government-sponsored projects (e.g., the Schizophrenia Patient Outcomes Research Team) (Lehman, Steinwachs, & PORT Co-Investigators, 1998), discipline-based groups (e.g., the American Psychiatric Association) (McEvoy, Scheifler, & Frances, 1999), and professional organizations for provider groups (e.g., the International Association of Psychosocial Rehabilitation Services) (Giesler & Hodge, 1998; IAPSRS, 1997b). Although practice guidelines differ widely in their level of specificity, the recent trend has been towards greater specificity so that adherence to guidelines can be measured. Fidelity measures and practice guidelines have not been explicitly combined as yet. However, it should be possible to develop fidelity measures that build on and extend practice guidelines.

Program standards

Program standards refer to a set of prescriptive program elements expected by some accrediting body. The marriage of program standards and fidelity scales is starting to emerge, especially with the ACT model. Specifically, the Commission for Accreditation of Rehabilitation Facilities has recently published a manual that includes a set of criteria for what constitutes an ACT program (CARF, 2000). These standards resemble a fidelity measure. Similarly, the Health Care Financing Administration is in the process of promulgating standards for approved ACT programs that will determine if a program can get reimbursed by Medicaid.

Program certification

Another approach related to the development of fidelity measures has been the desire of proponents of well-established approaches to maintain high standards as their model is disseminated.

An example is given by the Fountain House clubhouse model (Beard, Propst, & Malamud, 1982). Although the roots of this model date to the 1940s, it has evolved over the years and, like other psychiatric rehabilitation models, has yielded many variants. Recently, a set of clubhouse standards was adopted by an international conference of clubhouses, partly as a reaction to model diffusion (Propst, 1992). These standards have been the basis for a certification process, which is operated by the International Center for Clubhouse Development (ICCD), located at Fountain House. The formal process of certification is based on a site visit by a group of approved clubhouse trainers, who use a semi-structured guide in determining adherence to standards (Gold Award, 1999; Moxley, 1993). Clubhouses, therefore, are classified into "certified" programs (i.e., those approved by the ICCD), and "noncertified" programs. The ICCD site visit process has provided a foundation for a set of clubhouse fidelity indicators (Wang, Macias, & Jackson, 1999).

Report cards

Still another trend related to the interest in fidelity measures has been the role of consumer organizations in advocating for quality services. In state-by-state surveys, Torrey and his colleagues rated the adequacy of mental health services for people with SMI in the U.S., broadly characterizing domains such as case management and rehabilitation services in each state (Torrey, Erdman, Wolfe, & Flynn, 1990). The resulting state ranking provided a "report card" that consumers and families could use to evaluate services in their state and to advocate for better services. The National Alliance of the Mentally Ill has used a similar checklist approach to rate managed care organizations (Hall, Edgar, & Flynn, 1997). These evaluations suggest potential practical uses of fidelity measures for consumers. It remains to be seen how well the instruments developed by researchers can be adapted for such purposes. For example, many consumer report cards emphasize customer satisfaction dimensions, e.g., promptness of response and friendliness of staff, which comprise a much more general set of concerns than program specific fidelity measures.

General-purpose process measures

In the psychotherapy literature, a large "process" literature has accumulated over the past 3 decades, directed at uncovering the critical ingredients of successful psychotherapy independent of any specific program model (Orlinsky, Grawe, & Parks, 1994). Thus, much of this research has been atheoretical; working on the premise that a battery of process variables correlated with a set of client outcome variables will yield a core set of critical ingredients transcending any specific psychotherapeutic model. This inductive research strategy differs in its approach to identifying critical ingredients from theoretical (model-driven) approaches used

in fidelity measurement development. Both inductive and model-based research strategies have proven to be productive in the psychotherapy literature.

In like fashion, some psychiatric rehabilitation researchers have developed general-purpose instruments to measure program features across a broad range of programs (Burt, Duke, & Hargreaves, 1998; Jerrell & Hargreaves, 1991; Moos, 1974a). General-purpose process measures differ from fidelity measures by not explicitly postulating desired program characteristics. The most popular of these has been the Community Program Philosophy Scale (CPPS) (Jerrell & Hargreaves, 1991). The CPPS measures such dimensions as out-of-office contact, housing, links to entitlements, and longitudinality of care (Hargreaves, Shumway, Hu, & Cuffel, 1998), p. 110. The Community Oriented Program Environment Scale (Moos, 1974a; Moos, 1974b) is a popular process scale measuring the social environment, including dimensions such as autonomy, practical orientation, and anger and aggression. The COPES and related instruments developed by Moos have been widely used in psychiatric rehabilitation programs (Ryan, Bell, & Metcalf, 1982). More recently, Burt, Duke, and Hargreaves (1998) developed the Program Environment Scale, a questionnaire completed by consumers and staff in clubhouses and day programs. Preliminary data suggest that this multi-scale instrument discriminates well among different program approaches. The role of these generic instruments in fidelity measurement should be investigated further, perhaps as a supplement to more model-specific fidelity instruments (Brekke & Test, 1992).

The Promise of Fidelity Measures: Research and Practical Applications

The application of fidelity measures can be conceptualized along four dimensions: (a) purpose, (b) sample, (c) timing, and (d) target audience. With regard to purpose, fidelity measures can be used in research (e.g., ensuring model adherence in evaluations) or practice (e.g., monitoring program development). With regard to sample, fidelity measures can be used with a single program or a sample of programs. With regard to timing, fidelity measures can be introduced before an organization has even decided what program models might be implemented, at the initial stages of development of a new program, or at any stage after implementation. Fidelity measurement can take place once or at multiple time points. With regard to target audience, many stakeholder groups, including funding agencies, researchers, program managers and other program staff, consumers, and their families, are interested in maintaining standards for the sake of attaining quality of care. Applications of fidelity measures can, of course, incorporate several purposes, samples, points in time, and target audiences. We describe several different purposes for research and practice.

Research Applications

Fidelity measures have many applications in research (Moncher & Prinz, 1991). We describe four uses: (a) Ensuring model adherence in program evaluations, (b) Facilitating communication in the literature, (c) Synthesizing a body of research, and (d) Identifying critical ingredients of program models. We illustrate each of these uses with examples from the psychiatric rehabilitation literature.

Ensuring model adherence in program evaluations

Moncher and Prinz (1991) define fidelity of treatment in outcome research as "confirmation that the manipulation of the independent variable occurred as planned" (p. 247). This definition indicates one primary function of fidelity measures, namely, the importance of establishing the internal validity of a study. This is

done by showing that the independent variable has been successfully manipulated as planned and that this manipulation distinguishes the experimental from the comparison or control intervention in the desired way. Unlike classic experimental studies, in which one independent variable is manipulated or an active treatment is compared to a placebo, studies evaluating the effectiveness of a program model usually must contend with comparisons between two or more treatments, often with several "active" components. Thus, it is more difficult, but also perhaps more important, to demonstrate that the interventions did differ in expected ways.

Measuring program fidelity, then, is similar to conducting a "manipulation check," intended to determine if the independent variable yielded the desired difference between the treatment groups in the interventions they received. Making sure the interventions differ is also a way to increase statistical power (Lipsey, 1990). In order to test for significant outcome differences, the designs for program evaluations should "maximize the systematic variance" in the interventions provided (Kerlinger, 1986) by ensuring that the experimental and control groups are different. More generally speaking, it is important to examine the implementation of both the experimental and control groups along the same study dimensions in order to determine the degree of treatment differentiation, which is the systematic variance that is expected to account for any differences in outcomes. For example, one study contrasted two vocational approaches, which were sharply differentiated on the job placement process (initial prevocational skills training versus rapid job search) and on the location of services (at a separate rehabilitation agency versus at the mental health center where study participants received case management and other mental health services) (Drake, McHugo, Becker, Anthony, & Clark, 1996).

Measurement of fidelity is especially important in multi-site studies, in which findings may be stronger in some sites than in others. Program drift, that is, the tendency for programs to depart from the models that they are intended to replicate, is unfortunately quite common in the literature (Bond, 1991; Rosenheck, Neale, Leaf, Milstein, & Frisman, 1995). Multi-site randomized controlled trials that have used fidelity measures have been far more compelling in interpreting outcome differences from different sites (Drake et al., 1996; McHugo, Drake, Teague, & Xie, 1999). A related problem occurs when the control group starts to imitate the experimental intervention. Because of treatment contamination of control groups, it is important to measure fidelity in both the experimental and control conditions in a study. The goal in fidelity measurement in an evaluation study is not only to document degree of fidelity measurement, but also to provide information for making changes, if necessary. Rather than simply documenting failures to implement, a far more useful approach is to use fidelity measures in conjunction with systematic efforts to achieve excellence in program implementation, to help guide development, and to keep programs on course (Henggeler et al., 1999).

A key issue in ensuring program fidelity is to make sure that the program model is clearly defined at the outset. Initial ambiguity about the program model may have been a factor in the ambiguous findings in two recent large-scale studies (Becker, Holloway, McCrone, & Thornicroft, 1998b; Burns et al., 1999). In both studies, intensive case management was defined largely in terms of lower caseload ratios without detailed prescriptions for program implementation.

Another fidelity issue that is not frequently discussed in the literature is the need for continued fidelity measurement over the course of a study, especially for multi-year studies. Programs change over time, sometimes showing marked improvement as they move beyond their start-up period. Conversely, they sometimes lose their wholehearted commitment to the program model, for example, as they approach the end of grant fund-

ing, especially if this means (as it often does) that staff workers are transferred to other positions (McHugo et al., 1998).

Facilitating communication in the literature

One source of confusion in the psychiatric rehabilitation literature has been distinguishing between related program approaches within specific domains. Earlier, we gave the example of the case management area, in which many different variants have been reported in the literature. A recurring question has been the conceptual and empirical overlap between these different models (Mueser et al., 1998). Thus, one purpose for fidelity scales is to identify the distinguishing features of program models.

The Dartmouth ACT Scale (DACTS) is an example of a fidelity scale that has been helpful in mapping out a psychiatric rehabilitation domain (Teague et al., 1998). Although developed to discriminate well-implemented ACT programs from traditional case management services, the DACTS also may be useful for delineating a typology of case management in general. Another example of the use of fidelity measures for mapping out a domain is given by Bond, Becker, Drake, and Vogler (1997a), who found that vocational programs subscribing to the IPS model of supported employment sharply differed from "traditional" vocational services across a wide range of observable criteria. The IPS fidelity scale also found less dramatic, but significant, differences between IPS and other forms of supported employment.

One variation on the theme of using fidelity scales to map out a domain is defining model adaptations. Once a program model is well defined, variations of it can be assessed in relation to the original model. An example illustrating the use of a fidelity scale to measure model adaptation is given in a study examining the effectiveness of transferring clients from an ACT program to a modified ACT program referred to as a "step-down" program (Salyers, Masterton, Fekete, Picone, & Bond, 1998). This step-down program emulated many of the positive features of ACT, while serving clients at a less intensive level. The ratings on the DACTS indicated specific ways in which the step-down program differed from the parent ACT program (e.g., frequency of team meetings, percentage of home visits) and ways in which it was similar (e.g., use of multidisciplinary team, focus on practical problems).

Synthesizing a body of research

Literature reviews aim at understanding the extent to which findings from individual studies are generalizable, i.e., assessing external validity. In integrating the literature on the outcomes for a program model (or for an entire service domain, such as vocational services), reviewers face the dilemma of determining which studies to include and how to weight those that are included. Many factors go into these decisions, but one main consideration is the fidelity of implementation. The ideal circumstances for a reviewer would include a body of studies in which all investigators prospectively applied a single standardized fidelity measure, which previously had been shown to have adequate psychometric adequacy. Under these ideal circumstances, the reviewer could then establish a minimum criterion for program fidelity. If a program fell below the criterion, then that study would be excluded from the review. Alternatively, fidelity scores could be used as an independent variable in a meta-analysis (Lipsey, 1990).

Unfortunately, there are no examples in the mental health services area that come anywhere close to this ideal. Few domains have a critical mass of studies, and even fewer have used prospective fidelity ratings that would lend themselves to this procedure. However, two reviews using retrospective fidelity ratings hint at

the promise of this methodology. In one synthesis, McGrew et al. (1994) retrospectively coded 18 programs on a fidelity index, which was correlated with a program-level client outcome measure. This study found a strong correlation between the fidelity index and reduction in hospital use. More recently, Latimer (1999b) retrospectively coded programs within a sample of 34 ACT studies using a simplified fidelity scale and found that high fidelity ACT programs had better outcomes than low fidelity programs.

Identifying critical ingredients

A fourth use of fidelity scales is to help identify critical ingredients that predict client outcomes. Critical ingredients refer to the elements of a model, such as the caseload ratio or location of services, which account for its effectiveness. In this application, theoretically important ingredients are represented by items or subscales on a fidelity measure. The usual method for demonstrating empirically that a program element is a critical ingredient is by obtaining a significant correlation with a criterion measure (i.e., a measure of client outcome), controlling for other program characteristics. Following the logic of this design, the criterion measures should be congruent with the purposes of the program model; for example, a vocational model should have a primary impact in the employment domain. Although research has used different statistical methods, one typical strategy is to convert individual client outcomes into aggregate program-level measures (e.g., percent employed). This general strategy relates to the predictive validity of a fidelity measure by examining its relationship to outcome. Examples of this research strategy and limitations are discussed in Chapter 5.

Practical Applications

The potential practical applications of fidelity measures in psychiatric rehabilitation are numerous. They include communicating program standards, monitoring programs (over time and/or compared to other programs or norms), or to document the relationship between model adherence and outcomes. Many of these applications have already been attempted, although a complete inventory of actual applications has never been compiled. The intense interest in learning about and obtaining user-friendly fidelity checklists became apparent to us with the surprisingly enthusiastic response to an early publication of a fidelity instrument (McGrew et al., 1994). When fidelity is used in a practical setting, the results often go unrecognized in unpublished reports and documents. We illustrate practical applications of fidelity measures with the following examples.

Communicating standards

One practical use of fidelity scales is to introduce a program model to groups who have not had first-hand experience with it. For example, if a state mental health authority is seeking to introduce a new set of psychiatric rehabilitation services, it is helpful if decision-makers have concrete details before adopting a specific program model. Fidelity scales can provide a template for thinking about practice guidelines, whether or not a specific program model is adopted as is. Examples of task forces to develop practice guidelines can be found throughout the United States (Barton, 1997; Torrey & Wyzik, 1997), Canada (Cochrane, Durbin, & Goering, 1997; Latimer, 1999a) and overseas (Marshall & Creed, 2000). In addition to practice guidelines, fidelity measures can be used as a quick reference guide to program design and as a starting point for estimating program costs. Fidelity measures may also be used to communicate key elements of a program model to staff, consumers, and families.

MONITORING PROGRAMS

The history of deinstitutionalization has been a recurring cycle of reform movements, each one hoping to be the innovation that would take hold. Often with great fanfare, states announce initiatives to improve services, often through the introduction of a new program model.

New York's early experience with supported employment is illustrative of the implementation problems that occur as provider agencies implement wildly disparate services, often falling short of the effectiveness promised by the initiative (Noble, 1991).

Currently a number of states and local authorities are using simplified fidelity measures in the form of checklists as tools to help avert such mistakes. With appropriate databases, evaluators can provide cross-site monitoring of program implementation, making comparisons such as: (a) between target programs and established norms (if they exist), (b) across regions of the state (e.g., rural versus urban), (c) between individual sites and state averages, and (d) within programs and groups of programs over time. The comparisons can help identify specific areas in which the state as a whole falls short of established norms; regional differences that may be reflective of varying populations, resources, local traditions, or other factors; individual sites that may be exceptionally well-implemented and worthy of recognition, individual sites departing from the intended model and improvement over time as programs develop. From a management standpoint, it is valuable to know which sites are outliers, so that one can intervene early. Even a fairly crude fidelity measure may be capable of serving as an early warning system for such sites.

Program monitoring has been most widely used for ACT programs. Since the early 1980s, Michigan has sponsored the dissemination of ACT programs throughout the state, requiring new ACT teams to follow standards for program operation (Mowbray, Plum, & Masterton, 1998). Other states have followed suit (Deci, Santos, Hiott, Schoenwald, & Dias, 1995). Since 1996, Illinois has been monitoring agencies funded through their statewide ACT initiative (Zahrt, Bond, Salyers, & Teague, 1999). Using the staff from the state mental health authority to make ratings, program planners have found the DACTS to be a useful tool for communicating program expectations and ensuring their implementation.

One variation of a statewide monitoring approach concerns converting an existing program to a new program model. For example, Rhode Island recently has been involved in converting their day treatment services to supported employment (McCarthy, Thompson, & Olson, 1998). State planners in Rhode Island used the IPS fidelity scale to help shape expectations for provider agencies.

One of the most ambitious efforts in measuring program fidelity at a statewide level has been undertaken in Kansas. With the help of the University of Kansas School of Social Welfare, community mental health centers are currently using a comprehensive packet of materials known as "Best Practices Fidelity Tools" (Rapp, 1999). This packet spans the important domains of mental health services for adults with severe mental illness, providing specific behavioral indicators of what programs should be achieving in each service domain. Another example is given by Connect98, an initiative of the Illinois Office of Mental Health. This statewide program is using a set of fidelity tools to assess implementation of three psychiatric rehabilitation components: peer support, vocational services, and skills training (Bond, Evans, Kim, & Goodman, 1999b).

Aside from state-level monitoring, fidelity measures also may be useful for individual programs in a self-monitoring function. Programs may assess fidelity in their program over time to determine if significant changes

have been occurring in their implementation of the model. Similarly, if norms or other data are available, programs may compare themselves to others.

Documenting model adherence and outcomes. Another practical application is to use fidelity measures to document the relationship between program implementation and outcome. Although causal statements cannot be made without controlled research, descriptive information about program changes and changes in outcome can be useful. For example, programs can monitor fidelity and desired outcomes over time, noting changes that co-occur. A more sophisticated strategy might be to systematically change one aspect of the model (e.g., increasing adherence on an item on the fidelity scale) to see if changes in outcomes follow.

Discussion

We are seeing increasing demands for the measurement of adherence to program standards, not only from the research community, but also from a variety of stakeholders involved in funding, providing, and receiving psychiatric rehabilitation services. Increasingly, journal editors are insisting that empirical studies include fidelity measures. The National Institute of Mental Health, the Center for Mental Health Services, and other federal agencies increasingly are requiring fidelity measurement in grant applications to evaluate program models. Although not well documented, we believe fidelity measures also are being used more frequently in practical settings. The motivation for using fidelity measures is fundamentally the same for research and practical applications – to ensure that programs that say they are following a program model are in fact doing so. Clearly, the demand for these measures has grown in response to the problems that have emerged when fidelity has been ignored.

This chapter has hinted at the importance of fidelity scales for a wide range of audiences. Despite the fact that fidelity measures are now widely acknowledged as important for program evaluation, program implementation, accreditation, and financing, most fidelity measures in the psychiatric rehabilitation field are rudimentary. Careful psychometric work is needed if fidelity measures are to achieve their promise. We believe that we should apply the lessons from the psychotherapy fidelity literature, as well as the broader literature on measurement. Although fidelity measurement is no panacea, it can help in both the research and practice of psychiatric rehabilitation.

Chapter 2. Mapping Out the Domains of Psychiatric Rehabilitation

The term "psychiatric rehabilitation" has been used in many different, and often contradictory, ways (Anthony & Liberman, 1992; Beard et al., 1982; Cnaan, Blankertz, Messinger, & Gardner, 1988; Dincin, 1975; Liberman, 1988). Several authors have attempted to review the psychiatric rehabilitation literature. Limiting our attention to publications since 1986, these efforts include review articles (Baronet & Gerber, 1998; Barton, 1999; Dion & Anthony, 1987; Mueser, Drake, & Bond, 1997; Penn & Mueser, 1996) and books (Farkas & Anthony, 1989; Flexer & Solomon, 1993; Liberman, 1988; Pratt, Gill, Barrett, & Roberts, 1999; Stroul, 1986). In our view, psychiatric rehabilitation is not a unitary concept, but rather a set of ideas that include (a) fundamental values and attitudes toward mental illness, (b) specific goals to improve the lives of people with mental illness, and (c) a set of specific program models to accomplish these goals. One useful broad definition that fits most forms of psychiatric rehabilitation is offered by Rutman (1993): "giving people with psychiatric disabilities the opportunity to work, live in the community, and enjoy a social life, at their own pace, through planned experiences in a respectful, supportive, and realistic atmosphere." The literature suggests that psychiatric rehabilitation practitioners generally agree on many of the fundamental values and goals of psychiatric rehabilitation (Cnaan et al., 1988). However, there are conflicting viewpoints on the optimal service models to achieve these goals.

Psychiatric rehabilitation services are often compartmentalized into specific components, such as a specific vocational program. These components are often referred to as programs (or other names, such as service, department, etc.). Thus, for example, psychiatric rehabilitation agencies often have a vocational program, residential program, and a case management program as separate program components. Transcending particular program components, however, are a set of agency-level principles. In measuring fidelity of program implementation, fidelity scale users have focused on both principles of psychiatric rehabilitation (i.e., "agency-level) that transcend specific programs, as well as program models that refer to specific program components. Although this chapter will briefly review the broad agency-level principles, its main focus is on specific program models.

Psychiatric rehabilitation has evolved in an unusually eclectic and pragmatic fashion, because many areas lack clearly articulated models or theory explicitly linked to the nature of severe mental illness (Hogarty, 1995). In addition, many influential leaders in the psychiatric rehabilitation field have emphasized the advantages of eclecticism, innovation, and experimentation in program design (Dincin, 1995).

This chapter begins with a brief history of psychiatric rehabilitation, followed by a section on principles of psychiatric rehabilitation. Then, for each of several client domains (community integration, employment, independent living, social relationships, skills training, family relationship, and academic achievement), we identify a series of models reported in the literature. For each model, we describe the model and sketch some of its critical ingredients and summarize research evidence regarding its effectiveness. Finally, for each model, we provide information on fidelity measures available or under development. In the Appendix, we provide additional information on fidelity measures for each of the models discussed.

History of Psychiatric Rehabilitation

No discussion of psychiatric rehabilitation can begin without acknowledging the pioneering contributions of Bill Anthony. Starting in the 1970s with a series of literature reviews (Anthony, Buell, Sharratt, & Althoff, 1972; Anthony, Cohen, & Vitalo, 1978) and the first book devoted solely to psychiatric rehabilitation (Anthony, 1980), Anthony helped to provide a conceptual and theoretical rationale for what was, and still is, a large atheoretical and eclectic body of practices and philosophies. Over the last two decades, Anthony and his colleagues at Boston University have articulated their concepts and training approach. In their view, the foundations for psychiatric rehabilitation are rooted in the client-centered tradition (Carkhuff, 1969) and the skills training literature (Anthony, Cohen, & Cohen, 1984). They have devised a sequential model of practitioner intervention, starting with setting the overall rehabilitation goal, conducting a functional assessment, and then offering skills training. Anthony (1994) objects to calling his approach a "model," arguing that it is a set of procedures that should be incorporated into all program models. Although Anthony's early work emphasized the importance of model fidelity (Anthony et al., 1982), his psychiatric rehabilitation principles were disseminated worldwide (Farkas & Anthony, 1989) prior to extensive formal program evaluation.

Predating the work of Anthony was the development of the clubhouse model. It originated with the Fountain House program in New York City (Beard et al., 1982; Dincin, 1975). In the 1940s, the precursor to Fountain House was a self-help group for patients discharged from the state psychiatric hospital. The group sought a professional to serve as center director.

Under this new direction, the group evolved into an innovative program for helping clients with SMI adjust to community living. Operating outside the mental health system, the program became known as a clubhouse, because its identity revolved around a central meeting place for members to socialize. Propst (1992) outlined very prescriptive standards for the clubhouse including the following elements: membership, relationships, space, work-ordered day, transitional employment, independent employment, functions of the house, and funding, governance, and administration. Starting in the 1960s, Fountain House's success spawned a national network of independent psychosocial rehabilitation centers (Dincin, 1975).

Another important influence on psychiatric rehabilitation has been the consumer movement, which has stressed the importance of consumer involvement and empowerment. As noted in the previous chapter, the leadership of the office of the Community Support Program, which received its impetus and continued support through the National Institute of Mental Health (NIMH) and later the Center for Mental Health in the Substance Abuse and Mental Health Services Administration (SAMHSA), has helped shape the field of psychiatric rehabilitation. Many of the concepts of CSP overlap with and are consistent with psychiatric rehabilitation. Over a decade ago, Stroul (1986) compiled what at that time was the state of knowledge about CSP models, including a section on psychiatric rehabilitation, the clubhouse model, and various other program models.

Fundamental Concepts in Psychiatric Rehabilitation

Any discussion of the core ingredients of specific psychiatric rehabilitation models should be prefaced by mention of common ingredients of psychiatric rehabilitation in general. Using the Waltz et al. (1993) terminology, these are program elements that are essential but not unique. As is true for all of rehabilitation

psychology, psychiatric rehabilitation stresses the importance of individualizing rehabilitation planning, assessment, and intervention, and the focus on client strengths and the therapeutic relationship (Rapp, 1998). Thus, many of the specific principles enunciated below also apply to rehabilitation approaches for other client groups. Despite the diversity of psychiatric rehabilitation approaches, most psychiatric rehabilitation practitioners subscribe to a common set of guiding principles (Cnaan, Blankertz, Messinger, & Gardner, 1990). A partial list of these principles includes the following:

Pragmatism

Psychiatric rehabilitation has a focus on practical problems in everyday living (Dincin, 1975). Closely related to this pragmatism is an outcome orientation, wherein services are organized according to specific, tangible goals.

Attention to Client Preferences

Psychiatric rehabilitation programs attend to client preferences, for example, helping clients to find jobs in the occupations they desire (Becker, Drake, Farabaugh, & Bond, 1996) and to obtain their preferred types of housing (Carling, 1993).

Situational and Functional Assessment

Psychiatric rehabilitation programs generally use hands-on approaches to assess client abilities. Situational assessments -- observing clients in real-life situations -- often provide more useful information than standardized paper-and-pencil tests of ability (Anthony & Jansen, 1984). Functional assessments -- defining skill deficits that need to be addressed to achieve behavioral goals -- are used in some psychiatric rehabilitation approaches (Farkas, Cohen, & Nemec, 1988).

Environmental Modification

Psychiatric rehabilitation approaches emphasize the importance of selecting and changing environments to maximize the likelihood that clients will succeed. For example, staff may place a client with poor hygiene who is looking for work in a recycling center, where hygiene is less critical, rather than attempting to modify deeply-ingrained habits (Becker & Drake, 1993).

Integration of Services

Because traditional practice that separates vocational rehabilitation services from mental health treatment leads to fragmented services and poor employment outcomes (Noble, Honberg, Hall, & Flynn, 1997), many experts agree with Stein and Test (1980), that interventions are most effective when rehabilitation services are closely coordinated with treatment interventions.

Continuity of Services

The importance of maintaining continuity of services over time is another fundamental principle of psychiatric rehabilitation (Test, 1979). Because SMI involves chronic conditions, time-limited interventions are generally ineffective. Maintaining continuity in relationships by providing timely and predictable support is a key element in successful psychiatric rehabilitation programs.

Community Integration

Psychiatric rehabilitation embraces the principle of normalization, helping clients to move out of patient roles, treatment centers, segregated housing arrangements, and sheltered work, and enabling them to move toward illness self-management and normal adult roles in their communities. Studies by Drake et al. (1998) have shown that isolating day treatment programs can be closed down and replaced with programs to help clients find community jobs, with positive outcomes for clients, families, and mental health staff.

The Complications of Defining Psychiatric Rehabilitation

Psychiatric rehabilitation is part of a broad array of mental health services that are essential to the adequate care of people with SMI. Complicating the picture are services that many do not agree whether they fit under the rubric of psychiatric rehabilitation. Like others, we differentiate rehabilitation from treatments that are clearly part of a medical model, such as psychiatric medications, as well as psychosocial interventions that also are construed as part of the medical model, such as psychiatric hospitalizations, outpatient counseling, day treatment, and partial hospitalization. However, some psychiatric rehabilitation approaches sharply distinguish their services from medical treatments. The outstanding example of this attitude is found in the tenets of the clubhouse model (Beard et al., 1982). Other approaches take the opposite viewpoint, that helping people with SMI requires a holistic approach in which treatment and rehabilitation services are closely integrated (Test, 1992). Although psychiatric rehabilitation programs are often closely integrated with mental health treatment services, including medications and psychotherapy, these treatments are not included in the definition of psychiatric rehabilitation and are not discussed in this toolkit.

Case Management

Case management grew, in part, out of the return of consumers to the community after a period of deinstitutionalization in the 1950s and 1960s. This called for a need for the coordination of the many segregated services in the community. Several models of case management have been identified in the literature and have been in practice in the community. We briefly discuss four popular models: traditional case management, assertive community treatment, intensive case management, and the strengths model. Two other models (clinical case management and rehabilitation case management) are also reported in the literature, but have not been widely studied and will not be discussed here.

Traditional Case Management

Description

Traditional case management (TCM) refers to "standard" or "usual" services for clients, as provided by CMHCs and other service providers. It is no longer easy to give a clear definition of this approach, given the sea of change in the service system introduced by managed care. Historically, however, TCM has been defined as predominantly office-based services provided by bachelor's-level case managers who carry caseloads of 30 clients and often more (Boyer & Bond, 1999; Ellison, Rogers, Sciarappa, Cohen, & Forbess, 1995). Most often, TCMs carry individual caseloads, focus on entitlements and housing, and "broker" most of their assistance, that is, arrange appointments for their clients with other agencies or other staff workers who do much

of the direct services. TCMs may broker out services including day treatment, housing services, outpatient counseling, or vocational rehabilitation.

Effectiveness

It has been fashionable to criticize TCM, and it is certainly true that TCM, when provided to clients in crisis, has failed miserably (Latimer, 1999b; Mueser et al., 1998). Some research has shown that titrating the caseload size to client level of functioning is possible and desirable (Ryan, Sherman, & Bogart, 1997; Salyers et al., 1998).

Fidelity measures

There are currently no fidelity measures in use for TCM.

Assertive Community Treatment (ACT)

Description

Originally called Program for Assertive Community Treatment, the ACT model is based on six well-defined tenets: low client:staff ratio, provision of service in the community (e.g., in client's home), caseload sharing across team staff, 24-hour availability, provision of all services by the ACT team, and time-unlimited service provision.

Effectiveness

There have been many reviews of the ACT literature (Latimer, 1999b; Mueser et al., 1998; Rapp, 1998). These reviews have shown that ACT is very effective in reducing hospital use and increasing independent living and moderately effective in reducing symptoms and improving quality of life.

Fidelity measures

There are numerous fidelity scales for ACT currently being used in research and practical applications. The Dartmouth Assertive Community Treatment Scale (Teague et al., 1998) and the Latimer ACT Fidelity Scale (Latimer, 1999b) are described in the Appendix.

Intensive Case Management (ICM)

Description

ICM is not consistently described in the literature. Recently a study to identify the critical ingredients of ICM using experts was conducted by Schaedle and Epstein (2000). ICM is similar to ACT in many critical ingredients, including the emphasis on treatment in the community and a low client to staff ratio as does ACT. One of the main differences between the models is that intensive case managers do not share caseloads as they do in the ACT approach.

Effectiveness

The available research on ICM suggests that it helps to reduce unnecessary hospital use (Mueser et al., 1998; Schaedle & Epstein, 2000).

Fidelity measures

Although there are currently no published ICM fidelity scales, Schaedle (2000) has completed preliminary steps for constructing such a scale for New York ICM programs.

Strengths Model

Description

The Strengths model of case management has generally been described in terms of broad values and principles. This model, as the name suggests, focuses on the client's strengths and potentials rather than on their limitations. Rapp (1993) outlined the components of the strengths model: 1) the relationship between case manager and the client is essential and primary, 2) the focus is on client strengths, not pathology, 3) client determination is the impetus for treatment, 4) community resources are maximized, 5) client/staff interaction takes place in the community, and 6) the belief that people with mental illness can continue to learn, grow, and change.

Effectiveness

The strength model has not been widely studied, with only a handful of quantitative studies reported in the literature. Rapp (1998) has synthesized the larger case management literature, suggesting a set of critical ingredients common to effective case management, including the Strengths approach.

Fidelity measures

Rapp (1999) has recently developed a fidelity scale for measuring adherence to the strengths model.

Vocational Rehabilitation

Many different vocational models have been used to help people with SMI gain employment. We do not attempt to describe all of these here, but rather, highlight four approaches: the clubhouse model, diversified placement approaches, the job club, and supported employment.

Clubhouse Model

Description

As noted above, the clubhouse is a comprehensive psychiatric rehabilitation approach providing help in many domains of functioning and has been extensively described in the literature (Propst, 1992). The clubhouse model has been especially influential in the area of employment. Fountain House pioneered two key vocational concepts: the work-ordered day and transitional employment.

The work-ordered day involves members working in unpaid prevocational work crews on-site at the clubhouse (e.g., preparing noonday meals for members, answering the phone, cleaning the building) (Beard et al., 1982). Beard and his colleagues hypothesized that members benefited from doing chores necessary for the functioning of the clubhouse, because members would therefore feel needed for the success of the clubhouse. Transitional employment (TE) consists of temporary, part-time community jobs commensurate with members' stamina and stress tolerance designed to acclimate members to work and increase their self-confidence (Macias, Kinney, & Rodican, 1995). Clubhouse staff workers negotiate with community employers for TE

positions, which are typically entry-level jobs in a wide variety of settings. Members may hold several TEs and often return to the work units between paid jobs. The ultimate goal of clubhouse vocational programs is to help members achieve "independent employment" (that is, permanent competitive employment).

Effectiveness

Very little rigorous research has been conducted on the effectiveness of clubhouse model, although some surveys suggest employment rates of 40% or more for clubhouse members (Bond & Resnick, 2000). Recent work includes uncontrolled studies of the impact of clubhouse programs on employment (Macias et al., 1995), quality of life (Rosenfield & Neese-Todd, 1993), and social networks (Beard, 1992).

Fidelity measures

Lucca (1998) developed the Clubhouse Index to measure adherence to the Clubhouse vocational model. Although there are no other fidelity scales developed specifically for measuring clubhouse characteristics, there are a few examples (described in the Appendix) of implementation scales that have been used to assess the clubhouse environment (Burt et al., 1998; Macias & Jackson, 1993).

Diversified Placement Approach

Description

The term, "diversified placement approaches," has been used as a label to include a group of psychiatric rehabilitation approaches that, although they adhere to most clubhouse values (e.g., client empowerment, offering less demanding options before entering competitive employment, providing members a safe haven between temporary jobs), depart from the clubhouse standards in the way they conceptualize vocational services (Bond et al., 1999a). Diversified placement approaches differ from the clubhouse model by de-emphasizing transitional employment as the centerpiece of their strategy. Members are placed in a variety of paid jobs that are often supervised or protected jobs (i.e., not completely independent employment) without any time limits. Diversified placement approaches often capitalize on corporate employment initiatives to hire and train people with disabilities. Job development is redefined by forging business partnerships with owners, supervisors, or other corporate contacts who become inside advocates for the rehabilitation program. Agencies using diversified placement approaches include Thresholds in Chicago (Dincin, 1995) and the Village in Long Beach, California (Chandler, Levin, & Barry, 1999), although these two agencies differ substantially in many specific programmatic details.

Systematic research is needed to operationally define the approach outlined here and to describe variations in implementation. One dimension on which there may be variability is the extent to which programs require stepwise progression through intermediate levels of employment. A focus group consisting of Thresholds vocational staff identified four program dimensions: Prevocational (goal-setting, job preparation and skills training, money management, and self-reliance), Direct Vocational Services (assessing and developing jobs, job coaching, job site liaison, case worker responsibilities, work logistics, and ongoing support), Administration (marketing, staff training, paperwork), and Psychosocial (psychiatric issues, family issues) (Trochim, Cook, & Setze, 1994). This preliminary work may serve as a framework for developing more explicit standards.

Effectiveness

Although program evaluations on this approach have been encouraging, it has not been systematically evaluated in controlled research (Bond et al., 1999a).

Fidelity measures

Rollins (2000) have developed a fidelity scale specific to Thresholds, in conjunction with a randomized controlled trial comparing DPA to the IPS model. The scale in this case is a good example of the creation of a scale to assess fidelity to an approach that has been previously considered a formal model.

Supported Employment

Description

Among the various supported employment approaches, the IPS model (Drake et al., 1996) is the best defined and the most extensively studied. Other models of supported employment for people with SMI can be thought of variations on the IPS model. IPS is based on the following principles (Bond, 1998):

- Competitive Employment as Goal: The goal is competitive employment in work settings integrated in a community's economy.

Rapid Job Search: Consumers are expected to obtain jobs directly, rather than following lengthy pre-employment training.

- Integration of Rehabilitation and Mental Health: Rehabilitation is an integral component of mental health treatment, rather than a separate service.
- Attention to Consumer Preferences: Services are based on consumers' preferences and choices, rather than on providers' judgments.
- Continuous and Comprehensive Assessment: Assessment is continuous and based in real work experiences, starting from initial contact with consumers and continuing after a consumer is employed.
- Time-Unlimited Support: Follow-along supports are continued indefinitely.

Effectiveness

Bond, Drake, Mueser, and Becker (1997b) have summarized the research on supported employment for people with SMI. Overall, they found a mean competitive employment rate of 58% for supported employment clients, compared to 21% for clients in control groups receiving traditional vocational assistance, over follow-up periods of typically one year. Other indicators of vocational success (such as earnings from employment and job tenure) also favored supported employment. Several studies have shown that day treatment programs can be closed down and replaced with supported employment, leading to better employment outcomes and greater community integration, with no observed negative outcomes (Drake et al., 1994).

Fidelity measures

We list several fidelity scales in the Appendix that have been developed for measuring the adherence to various supported employment models, including IPS.

Job Club

Description

The job club is a structured behavioral approach to help unemployed persons find jobs (Azrin & Philip, 1979). It provides systematic guidance in developing job leads, making telephone and in-person contacts, and obtaining jobs. Methods have been refined for teaching skills for interviewing for a job. In addition, job clubs use peer support as a way to encourage clients to continue the search process. In many applications, the job club is a "stand-alone" program, with the assumption that providing support and teaching skills will enable clients to find jobs on their own. Because of the availability of a program manual (Azrin & Besalel, 1980), it is readily amenable to a fidelity scale development.

The job club has been adapted for use in the psychiatric population (Jacobs, Kardashian, Kreinbring, Ponder, & Simpson, 1984). Jacobs et al. (1984) concluded that for persons with SMI, the job club needed to be adapted to provide more direction, interpersonal support, and encouragement from counselors than in the standard approach.

Effectiveness

Some studies have suggested that the job club has encouraging rates of job acquisition with some groups of clients with SMI (See Bond, Drake, & Becker, 1998a). However, job clubs apparently are most suitable for people who already have adequate social skills, especially interviewing skills. Dropout rates are high for people with severe disabilities (Corrigan, Reedy, Thadani, & Ganet, 1995; Jacobs, Wissusik, Collier, Stackman, & Burkeman, 1992).

Fidelity measures

We do not know of a published fidelity scales for measuring adherence to the job club, although this model would be amenable to such assessment.

Other Domains

Supported Education

Description

Educational programs are one option for higher functioning clients, who may not be challenged sufficiently by traditional psychiatric rehabilitation programs (Hatfield, 1989). Supported education helps clients obtain education and training in order to have the skills and credentials necessary for obtaining jobs with career potential (Moxley, Mowbray, & Brown, 1993). Unger, Danley, Kohn, and Hutchinson (1987) were among the first to pilot the concept of supported education. Unger (1998) has recently synthesized many of these ideas into a monograph. Starting in the late 1980s, the supported education concept was applied specifically to training clients to work as mental health paraprofessionals (Sherman & Porter, 1991). This idea has been widely emulated (Mowbray, Moxley, Jasper, & Howell, 1997).

Effectiveness

Although there are a few rudimentary studies on the effectiveness of supported education efforts (See Unger, 1998), to our knowledge, no randomized controlled trials have yet been conducted.

Fidelity measures

We are not aware of any fidelity scales for supported education.

Skills Training

Description

Although skills training typically focuses on social skills, it also has been used for a range of other skills needed for independent living. The goal of social skills training is to systematically teach the component skills necessary for effective social interactions. Typically, the steps in skills training are as follows: (1) give a rationale for learning the skill, (2) role-play the skill, (3) provide an exercise in which the client role-plays the skill, (4) give specific positive and corrective feedback on the client's role play, (5) have the client practice the skill, and (6) give a homework assignment in a real-life situation (Mueser et al., 1997). Many skills training curricula have been developed, including a widely disseminated set of videotape-guided modules (Liberman, 1985).

Effectiveness

Skills training has been the focus of many effectiveness studies. In an early review, Dion and Anthony (1987) report that most studies of social skills training were effective in improving social skill deficits in psychiatric patients. In a more recent meta-analysis, Dilk and Bond (1996) reviewed 68 studies looking at the effectiveness of skills training. The authors conclude that skills training is effective in acquiring the taught skills, particularly in inpatient settings (where most of the studies were conducted). However, the authors emphasize that it is still unknown whether these learned skills transfer to other settings such as the community or outpatient settings.

Fidelity measures

Wallace et al. (1992) created a Therapist Fidelity Checklist to measure how well leaders of skills training groups were adhering to the outlined criteria. This scale is described in the Appendix.

Drop-in Centers

Description

Peer support and socialization, although an important part of psychiatric rehabilitation, are less well defined than other areas. This type of rehabilitation can take place in many forms through relationships with staff, other members, and the outside community. There are several service initiatives that subscribe to the peer support paradigm. These include consumer-run businesses, advocacy organizations, consumers as case managers and other mental health staff, and drop-in centers (DIC). The DIC is by far the most common.. In the 1970s, Webb (1973) described the DIC as helping peers to facilitate socialization for each other, suggesting that goals of reduced hospitalization, increased socialization, and community integration could be achieved through intense member involvement in the organization and development of the program. The DIC is typically a place where consumers can come to socialize, obtain information or assistance from staff and other members, and generally a place where they can feel welcome and obtain a sense of belonging. An important aspect of the DIC is that it is consumer-run, that is, consumers feel a sense of ownership of the program. Beyond these basic characteristics, there is no formal structure for the DIC.

Effectiveness

Although there have been a smattering of outcomes studies relevant to drop-in centers, most have been either uncontrolled or used drop-in centers as control groups (Bond et al., 1995; Bond et al., 1990; Morse, Calsyn, Allen, Tempelhoff, & Smith, 1992; Mowbray, Chamberlain, Jennings, & Reed, 1988; Mowbray & Tan, 1992; Silverman, Blank, & Taylor, 1997). A multi-site study funded by SAMHSA is currently under way to study the effectiveness of consumer-run DICs.

Fidelity measures

Although there are no scales that have been developed that measure the DIC model broadly, several scale have been used to measure the environment of these types of programs (Evans, Resnick, & Bond, 1998; Macias & Jackson, 1993; Mowbray, 1999).

Housing

Description

There is no simple way to describe rehabilitation models in the housing area. Although supported housing is probably the most commonly discussed housing option in the psychiatric rehabilitation field, a wide range of other housing options are also often used. These include group homes (with different levels of staff involvement), apartment buildings managed by psychiatric rehabilitation agencies, and scattered-site housing, to name a few (Dincin, 1988). The terminology used for the various housing options is unstandardized.

In the 1960s, the prevailing philosophy for helping psychiatric patients return to the community was training them in a gradual, stepwise fashion to gain the skills to function in normal society. The paradigm was that discharged patients were first transferred to supervised group homes ("halfway houses"), later to a less supervised setting, and eventually to independent housing. One early transitional housing approach was the lodge model (Fairweather, Sanders, Cressler, & Maynard, 1969). Originally designed as an approach to the transition from psychiatric hospitals, most lodges now in operation are more akin to permanent congregate housing programs. Fairweather created a program manual and guidelines long before they were fashionable in the psychiatric rehabilitation field. They are self-contained societies in which persons with SMI live, work, and socialize together. A directory compiled a decade ago listed 100 lodges (Fergus & Balzell, 1990). The lodge model never achieved the national expansion Fairweather had thought possible (Fairweather, 1980), in part because deinstitutionalization made some elements of the model obsolete.

As mentioned above, one prominent philosophy of housing for people with SMI is supported housing, which involves living in independent community housing, with sufficient housing subsidies available to enhance the affordability of housing choices. Independent living arrangements are based on consumer preferences (for type of housing, roommates, location, etc.), and augmented with the appropriate level of professional and nonprofessional support necessary for the individual to manage independent living skills (Carling, 1993). Rapp (1999) has developed a fidelity scale for supported housing that captures these and other elements.

Effectiveness

Although adequate housing is widely regarded as a critical element in the recovery process, there are no recent reviews we know of that comprehensively review this area.

Fidelity measures

We list two scales in the Appendix that have been developed to measure adherence to residential program models.

Family Psychoeducation

Description

Family psychoeducational interventions are highly structured approaches that aim to lower the emotional climate in the family and to help family members manage the relative's mental illness (Lam, 1991). Different models of family psychoeducational interventions have been developed over the past two decades (Penn & Mueser, 1996). Behavioral Family Management (BFM; Falloon, 1984) is an approach conducted in the family's home based on behavioral therapy and social learning theory. In this model, a wide range of behavioral approaches is used to promote efficiency of family coping and communication skills. Modified versions of BFM (McFarlane et al., 1995; Randolph et al., 1994; Tarrier et al., 1989) include the use of role play and problem solving and may take place in multiple-family groups.

Broad-based psychoeducation (Hogarty et al., 1986; Leff, Kuipers, Berkowitz, & Sturgeon, 1985) is another commonly used model which aims to increase the stability of the home environment by promoting effective stress management. These models share common components to provide education about the nature and symptoms of schizophrenia and its treatment including medications and side effects, training in problem solving skills, and emotional support from clinicians or/and other peer families (Lam, 1991; Mueser & Glynn, 1995). In addition, these models share similarities such as establishment of a therapeutic alliance between clinician and family and an attempt to modify the home environment to minimize stress. These family psychoeducational interventions differ, however, in terms of format such as location, frequency, length and duration of session, inclusion of clients in the session, individual vs. multifamily session as well as techniques of training in communication and problem-solving skills (Halford & Hayes, 1991; Penn & Mueser, 1996).

Effectiveness

Controlled studies found that clients who received family psychoeducation reported significantly fewer relapses, more sustained remissions, and improved social functioning than clients in a control group. In addition, the intervention had a positive effect on lowering distress and the burden families experience, suggesting that it reduced emotional tension at home (See studies cited above). However, the process through which family psychoeducation works still needs to be studied. In a similar vein, the question of whether different models of family psychoeducation are equally beneficial to clients and families remains unanswered though several researchers have attempted to compare different models (e.g., Mueser, Gingerich, & Rosenthal, 1994; Tarrier et al., 1989; Zastowny, Lehman, Cole, & Kane, 1992).

Fidelity measures

Mueser and Glynn (1999) have developed the Therapist Fidelity and Competency Scale to measure adherence to their model. This scale is described in the Appendix.

Discussion

The most significant influence of this growing emphasis on fidelity measurement is the pressure to define program models operationally. Historically, psychiatric rehabilitation has been a field that has prided itself in its innovation, creativity, and flexibility. Not everyone is keen on the idea of program models. IAPSRS, for example, has pointed out an alarming trend for state and local mental health authorities to fall prey to the "single model trap" (Hughes & Clement, 1999; IAPSRS, 1997a). The implicit assumption with an emphasis on a single model is that "one shoe fits all." When policymakers stipulate that funding be provided only for a specific model of services, they foreclose other options. In response, providers often argue that their existing eclectic programs are equally or more effective for their clients, or that the designated model is inappropriate for some segments of their population. According to this reasoning, the idea of a single program model is antithetical to the idea of a "flexible array of options" that has been at the heart of the psychiatric rehabilitation philosophy.

Another historical objection to promulgating program models is given by Bachrach (1988), who argued that model programs are developed in a particular sociocultural and economic context that do not generalize to local conditions. Providers know their constituencies best, and they should adapt the models to their conditions, rather than mechanically apply what the original developers did. This particular argument has been applied repeatedly to the Madison model of ACT (Stein & Test, 1980) over the past 20 years.

We agree that these "anti-model" viewpoints have merit. Unfortunately, such arguments can be conveniently used to maintain the status quo, to reassure providers that whatever they are doing is adequate. Sometimes programs follow no apparent model at all, with individual clinicians varying widely in their intervention strategy. Another common response by providers is to blend an existing approach with the new model, accepting elements of the new model while discarding others. If the philosophy of operationally defining program models and then testing their effectiveness is applied to alternative approaches, then we see no conflict between the "model" viewpoint and the viewpoint of critics.

For measuring fidelity, we propose that the ideal way to build a fidelity measure is to begin with a model.

However, there are occasions when there is a need to measure the fidelity of a program for which there is no well-established model, for example, a relatively unique program such as a specific vocational approach that has evolved over time in a local community. In this case, there may be no suitable fidelity scales currently available. The need for such scales is especially apparent for investigators who are proposing a new set of interventions for a specific target group, such as people with schizophrenia and post-traumatic stress disorder or early intervention programs for young adults at risk for developing severe mental illness. Fidelity scales need to be developed and validated for many program areas, as suggested in the Appendix, which outlines fidelity scales in use. For example, scales are needed for supported housing, supported education, and consumer-run drop-in centers. Quite clearly, the most usual case is that an investigator will find that no scale exists that adequately measures the model program of interest.

Most importantly, studies that use fidelity measures and more general measures of program characteristics should be used to empirically address these issues. Such studies can tell us whether program models can be transported to different localities. They can also tell us how well different models, when implemented

with fidelity, work with different types of persons. Finally, they can tell us the degree to which program models, taken as a whole, or certain ingredients, common across models, are associated with impacts.

Chapter 3. Preparing for Scale Development

The next three chapters have been organized to mirror the sequential steps in developing a fidelity measure, as shown in Table 3.1. We will discuss Steps 1-4 in Chapter 3, Steps 5-11 in Chapter 4, and Steps 11-14 in Chapter 5. These steps are not unique to fidelity measurement, but are the appropriate steps for the development of any type of social or psychological scale (e.g., attitudinal, personality) (Hinkin, 1995; Nunnally & Bernstein, 1994). Fidelity measurement differs from most measures in that the level of analysis is the program, not the individual. Many of our examples of scale development and data collection methodology focus on the use of interviews, paralleling the popularity of this approach found throughout the current fidelity research.

Table 3.1 Steps for Developing a Fidelity Measure

1.	Define the purpose of the fidelity scale
2.	Assess the degree of model development
3.	Identify model dimensions
4.	Determine if appropriate fidelity scales already exist
5.	Formulate fidelity scale plan
6.	Develop items
7.	Develop response scale points
8.	Choose data collection sources and methods
9.	Determine item order
10.	Develop data collection protocol
11.	Train interviewers/raters
12.	Pilot the scale
13.	Assess psychometric properties
14.	Determine scoring and weighting of items

Step 1 Define the Purpose of the Fidelity Scale

The first step in developing a fidelity measure is to define its purpose. In Chapter 1, we described fidelity measurement as having both research and practice applications. We then defined specific goals within each of these sets of applications. The goals of a fidelity scale will influence the tactics used to develop the scale. For example, if the goal is to develop a scale for demonstrating model adherence in a randomized controlled trial, then the methods used will likely be more comprehensive, identifying features that make the model unique, and features that distinguish the model from services received by control groups. The evaluator is more likely to consider multiple measures, to conduct detailed reliability studies, and to administer the fidelity scale repeatedly. Conversely, if one is conducting a low-budget, statewide survey, where the goal is to ensure that sites achieve a minimal level of compliance to a program model, then a more pragmatic strategy is likely to be employed.

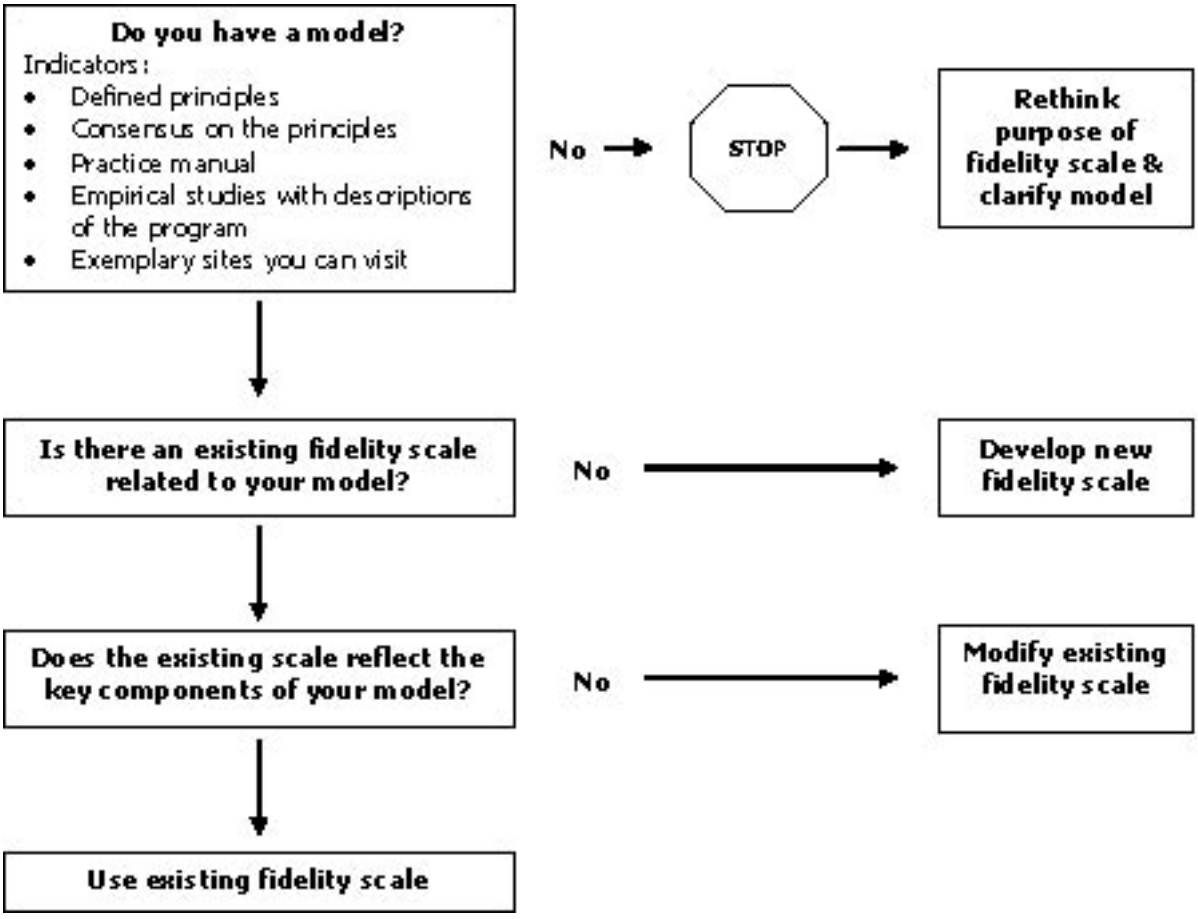
Step 2 Assess the Degree of Model Development

As suggested by Figure 3.1, the next step is to assess the degree of model development. If the program in question is well defined, then this suggests the use of confirmatory methods. If the program is not well-defined, then inductive methods may be more appropriate.

The assessment of the adequacy of a program model includes a literature review. First, review the literature on the particular program model to identify the important dimensions in the model as well as provide a more coherent understanding of the definitions of the constructs therein. (In this chapter, we use a variety of terms – principles, components, elements, and ingredients – to refer approximately to the same thing.) Second, the evaluator should review any existing literature on fidelity measures that have been designed for the particular program. This could help to determine whether there is an existing scale that can be used, or modified, or whether a new scale should be developed. The literature may also indicate particular dimensions that are difficult to assess or suggest which data sources are most appropriate (e.g., use of client self-report for a drop-in center).

A review of the literature will help to determine the degree of model clarity, model specification, model differentiation, model comprehensiveness, and model consensus. Model clarity refers to the extent to which the program model has clearly articulated principles of operation. An example of a program principle is "rapid job search." Model specification refers to the degree to which the model has explicit behavioral guidelines for operation. For example, the model specification for the principle of assertive outreach might be "at least 3 contacts per week at the consumer's home." Model differentiation refers to a distinctive feature of a program model that sets it apart it from other models and approaches. The use of a total team approach differentiates ACT from intensive case management. Model comprehensiveness refers to the extent to which a model provides adequate guidance for commonly occurring situations. Many theoretical models are inadequate by virtue of the fact that they do not tell what to do in important circumstances. For example, consider the fact that many case management models do not explain how to handle the management of the consumer's income. Model consensus refers to the degree of agreement with which publications in the field share a description of a model. "Clinical case management" is an example of a model lacking model consensus.

Figure 3.1 Decision Tree for Preparing for Fidelity Scale Development



Step 3 Identify Model Dimensions

Once it has been established that a model exists, the next step is to identify what the important elements of the model are. The elements must eventually be identified at the level of operationally defined items. In general, there are two broad strategies – confirmatory methods (when the elements of a program model are already well defined) and inductive methods (when the program approach has not yet been formally defined as a model). We will talk about these two strategies in turn.

Confirmatory Methods

We use the term confirmatory methods to refer to strategies used to formally identify program components for a program model when a model is already well described in the literature. Thus, confirmatory methods begin by documenting what is known in the literature about the program model, relying on evidence-based practice when this is available (Hughes, 1999). Papers that describe program principles (e.g., Bond, 1998; Propst, 1992; Rapp, 1998; Test, 1992; Witheridge, 1991) are very helpful in this regard. Practice manuals (e.g., Allness & Knoedler, 1998; Becker & Drake, 1993) are another useful source. Visiting programs representing exemplars of a model is another method of supplementing information on program elements.

We have used the broad categories of Staffing, Organization, and Services as a framework to begin generating items when developing fidelity measures (Bond et al., 1997a; McGrew et al., 1994; Teague et al., 1998). Although this framework does not correspond to the factor structure from empirical studies (Bond, 1999; Bond, Picone, Mauer, Fishbein, & Stout, 1999c; Teague et al., 1998), it does give the evaluator a place to start in the generation of items.

One issue in defining program dimensions is to ensure that the framework is comprehensive. To increase comprehensiveness, we suggest creating a grid mimicking the work of Waltz et al. (1993). Create a table with 5 columns. The first column should consist of theoretical dimensions (or critical ingredients). The next four columns should indicate if each dimension is (a) essential and unique, (b) essential, but not unique, (c) compatible, or (d) prohibited. The grid may help identify areas where the proposed conceptualization is incomplete. The fidelity measure should include essential dimensions, some of which may be unique to the particular model. If specific practices are prohibited in a model (e.g., prevocational work crew in an IPS program or recreational activities during working hours in a clubhouse), then it is often useful to include items reflecting these when assessing fidelity. Such items would of course be reverse-coded in a fidelity scale.

Often a program model already has well specified program guidelines. These guidelines could be used directly to indicate the program dimensions. Typically confirmatory methods employ a second step, which is to document the degree of consensus among experts, users (i.e., practitioners), and other relevant groups who are knowledgeable about the model. For shorthand, we will refer to such a group as an expert panel, although the composition of such group can be varied. For the step of identifying program dimensions, we recommend the use of an expert panel in most cases, even when the program elements are already well known.

Expert panels: Background

Expert panels have been used for a wide variety of purposes. Perhaps the most relevant literature for fidelity scale development is found in papers describing the development of practice guidelines in medicine (Brook, 1989; Eddy, 1990a; Eddy, 1990b; Woolf, 1990; Woolf, 1992). Although there are important differences in the development of practice guidelines and fidelity scales, the principles articulated in the aforementioned articles are relevant to the current context. In particular, Woolf (1992) distinguishes between informal and formal consensus development:

Informal consensus development refers to guidelines based solely on expert opinion. "Guidelines issued by specialty societies, federal agencies, and task forces have generally emerged from meetings of expert panels in which agreement is reached through open discussion, sometimes producing recommendations in a single meeting" (p. 946).

Formal consensus development was introduced in the U.S. in the 1970s. The National Institutes of Health developed a structured 3-day conference of experts involving group discussion following a specific format. Other formal consensus methods have involved mailings to experts and aggregating results, which Woolf (1992) notes is not a truly consensus process. In the 1980s, the RAND Corporation developed the currently most widely adopted expert panel approach. Like other expert panel approaches, the RAND methodology includes face-to-face meetings with a carefully chosen group of experts. One key innovation in the RAND methodology involves providing the panel with a systematic body of scientific evidence in preparation for the formal meetings. A two-step Delphi technique is then used. In the first step, panel members respond to

a list of medical procedures for the treatment of a specific medical condition. Panel members independently make ratings of appropriateness for each procedure. Next, the panel meets together as a group and formally discusses any disagreement. Woolf (1992) criticizes the RAND technique, which has become increasingly popular, because it "does not provide an explicit linkage between recommendations and the quality of the evidence" (p. 947).

The take-home message from this body of work for fidelity scale development is to acknowledge the limitations of clinical opinion and emphasize the importance of explicitly defining the methods used in any fidelity scale development.

Expert panels for identifying dimensions

None of the current work on fidelity scale development has approached the level of sophistication found in expert guideline development. However, the principles stated in the aforementioned articles are useful reminders of methods that will improve our fidelity measures. Key principles include the use of best evidence, making explicit the knowledge base in identifying the critical dimensions, and making explicit the steps used to reaching the final list of program elements.

Expert panels may be helpful in addressing two separate features of critical ingredients. These are: (1) Importance: What elements are essential, and (2) Model Specification: What level or intensity of the program element is viewed as critical. For example, an item on an expert survey might be, "How important (or essential, or critical) are treatment team meetings to the program model?" The companion question would be, "How often should the treatment team meet? For both types of questions, respondents should be given appropriate definitions, e.g., what is meant by a team meeting (do administrative meetings count?). For model specification, the respondent should be given an appropriate context, e.g., how many meetings per week and how many program staff for a program serving 50 clients.

Evaluators have used a variety of formats for obtaining opinions from panel members. For the confirmatory method, we recommend obtaining information independently from panel members rather than in a group format to avoid the influence of the group. Two ways to obtain information is via interviews and self-administered questionnaires. An example of an interview method is given in a study examining the ACT model (McGrew & Bond, 1995). Structured telephone interviews were conducted with 20 nationally-known experts on the ACT model, using the Critical Components of ACT Interview (CCACTI), a 73-item checklist adapted from checklists used by two state mental health agencies. Respondents rated each item on a 7-point scale of importance to determine the critical ingredients, as shown in Table 3.2. They also indicated the "model specifications" for elements they deemed important, as shown in Table 3.3. Two subsequent studies have adapted their methods from the McGrew and Bond (1995) study (Marty, Rapp, & Carlson, under review; Schaedle & Epstein, 2000). Marty et al. (under review) examined the critical components of the strengths model of case management adapting the CCACTI and using a similar interview format. One advance over the McGrew and Bond methodology was the use of two expert groups -- academically-based experts and practitioners -- allowing for comparisons in these two perspectives. Another improvement was that the researchers included distractor items in their inventory. That is, they included items they believed would not be highly endorsed, to demonstrate that respondents were not simply endorsing all items presented. Using a mail survey, Schaedle and Epstein (2000), also used an adapted version of the CCACTI to identify the key components of intensive case management (ICM). They used three distinct samples: researchers/administrators, program manag-

ers, and case managers. Following the completion of the self-administered checklists, Schaedle and Epstein conducted focus groups with case managers to interpret survey responses. This study demonstrated areas of both convergence and divergence among the respondent groups.

Table 3.2 Expert Ratings of Importance on Selected Items on CCACTI (adapted from McGrew & Bond, 1995)

Item	Mean (SD)	% "Very Important"
One team member is coordinator	6.9 (0.2)	90%
Coordinator responsibilities limited to ACT	6.7 (0.7)	78%
Shared caseloads for treatment planning	6.6 (0.9)	74%
Multidisciplinary team	6.6 (0.7)	72%
Psychiatrist on team	6.5 (1.1)	70%
Team size at least 3 FTE	6.5 (0.8)	69%
All team members attend all meetings	6.4 (0.9)	65%
Registered nurse on team	6.3 (1.1)	65%
Shared caseloads for service provision	6.1 (1.3)	58%
In vivo treatment focus	6.9 (0.3)	89%
Petty cash fund	6.5 (1.0)	72%
Answering machine/service	6.3 (1.7)	77%
24 hour availability	6.1 (1.3)	58%
NOTE: Rating scale: 7 = Very Important...1 = Not Important		

To identify a core set of critical ingredients (e.g., like those in Table 3.1), it is necessary to establish a criterion for item inclusion. There are no established rules for setting this criterion. McGrew and Bond (1995) arbitrarily used the criterion that at least 50% of the experts rated an item as "Very Important." Using this criterion, 54 of 73 items were rated as critical.

Identifying the critical ingredients does not indicate how much is needed of an ingredient. For example, experts may agree that a small caseload is critical, but they may differ on how small a caseload is needed. McGrew and Bond (1995) used the mean and standard deviations of model specifications reported by experts, as shown in Table 3.3, but other decision rules for defining model specifications also could be used (e.g., median, mode, range).

Table 3.3 Experts' Judgments of Ideal Model Specifications for ACT

(adapted from McGrew & Bond, 1995)

Model Specification	Mean (SD)
Maximum client:staff ratio	13.0 (2.7)
Optimal client:staff ratio	10.1 (2.6)
Mean number of contacts/wk	3.0 (1.5)
% of contact in home/community	75.3 (12.6)
Psychiatrist time(hrs/wk) (per 50 clients)	13.2 (8.8)
Nurse time(hrs/wk) (per 50 clients)	31.7 (9.8)
Maximum caseload size	97.7 (32.4)
Number of team meetings/wk	5.5 (1.9)
Length of team meetings(min.)	57.5 (15.0)
New clients admitted/month	6.4 (5.0)

Another recent expert panel produced the Expert Consensus Guideline Series for the Treatment of Schizophrenia, originally in 1996 (Frances, Docherty, & Kahn, 1996), and later updated (McEvoy et al., 1999). In the most recent version, McEvoy et al. (1999) used mail-out surveys to experts after creating an "algorithm based on the existing research literature and published guidelines" (p. 9). The 1999 task force created three written questionnaires concerning medication treatments, psychosocial treatments, and policy issues. The 68 experts for the psychosocial survey were identified from literature reviews and recommendations from professional organizations. The survey asked experts to rate a series of items on a 9-point rating scale used in earlier expert panels.

Our impression is that the expert panel method just described is susceptible to an inflated number of critical ingredients (Marty et al., under review; McGrew & Bond, 1995; Schaedle & Epstein, 2000). To put is simply, expert panels tend to rate most plausible items as "Very Important." (Who would be opposed, for example, to assertive outreach?) To counteract the tendency to over endorse items in closed-ended survey, we recommend that evaluators include distractor items (i.e., items that are either not essential or actually prohibited). Having some items that are not heavily endorsed makes the survey findings more credible. Another way to counteract the tendency to over endorse is to stipulate to panel members that only a certain quota of items can be given the highest rating (Drebing & Van Ormer, 1999). This latter method has its own drawbacks, as it may artificially exclude items that should be rated as critical. To our knowledge, there is no perfect solution to this dilemma.

Confirmatory approaches require that the investigators generate items a priori. However, even when using a confirmatory methodology, we recommend supplementing closed-ended surveys with open-ended questions. The intent is to identify any gaps in the original inventory of items. We should mention that our experience is that respondents usually do not volunteer many new items when the closed-ended part of the survey is extensive (McGrew & Bond, 1995). One potential strategy is reversing the order of inquiry (ask open-ended questions first), so as not to cue for the "right" answer. We do not know of any investigators who have so reversed the order of inquiry in an expert panel.

Another useful technique is to assemble a second panel to interpret findings from an initial expert panel survey. Schaedle and Epstein (2000) used this strategy very effectively, asking focus groups to interpret the expert survey responses. We recommend that evaluators consider including a supplemental method to help interpret closed-ended survey data.

A second critical decision concerns sampling, including both sample size and sample composition. Ideally, the sample size used should be derived from level of precision desired for estimates of the mean response, following well-known statistical principles. Previous studies can be useful in making this determination. When a model is well known, then there may be a high degree of consensus among experts on what are the critical dimensions. In these circumstances, a relatively small sample may be sufficient. For example, McGrew et al. (1995) interviewed only 20 experts in their ACT survey. This sample was sufficient for demonstrating a clear consensus on the majority of ACT items, although a larger sample would have permitted more fine-grained distinctions, as Meisler and Olsen pursued in a subsequent survey (Meisler, 1997). Our impression is that even when there is good consensus on the critical ingredients, experts often disagree on the specifics of model specification. For example, generally there is better agreement that an area is important (e.g., frequent contact with consumers) than what the specific guidelines should be (e.g., 2 contacts per week).When there is variability in opinion, larger samples are required to achieve desirable levels of precision.

The composition of the respondent sample will depend on the goals for the survey and the type of "backing" the evaluator wants for the fidelity scale. Some evaluators will conclude that an expert sample is most desirable. McGrew and Bond (1995) used an opportunity sample, selecting experts based on reputation and publications in the literature. Both Schaedle and Epstein (2000) and Marty et al. (under review) used at least two sampling methodologies, to address sample biases. We recommend that evaluators use multiple perspectives, documenting well the method for identifying respondents. McGrew and colleagues subsequently obtained information on critical ingredients of ACT from both clients (McGrew, Wilson, & Bond, 1996) and case managers (Test, Bond, McGrew, & Teague, 1997). However, these later studies did not use the identical format for obtaining perceptions as the original expert survey, making direct comparison difficult. Therefore, we recommend that, when obtaining information from multiple informants, evaluators should use the same procedures and the same format, when this is feasible.

Inductive Methods

Confirmatory methods are best for confirming what one already knows; they are not useful for discovering new elements not already part of the evaluator's preconceived notion of how a program model works. Inductive methods are used to "flesh out" or uncover the tacit elements in a program approach.

In this section, we consider the special case: What do you do when you don't have a model? Referring back to Figure 3.1, our general answer is, to stop scale development. However, we recognize there are times when evaluators seek to measure program implementation in the absence of a program model. The following discussion is devoted to the situation in which the evaluator is seeking to induce a program model, based on program practices. We also believe that inductive methods can be used to supplement confirmatory methods, that is, the situation when the program model is well known.

Carefully controlled program evaluation studies pose a situation in which one would try to define a program model when the users of an approach may not consider that their program service constitutes a formal model.

When the reason for developing a fidelity scale is to ensure treatment adherence in a randomized controlled trial, it is often the case that the comparison group in the study is not well defined. An example of is given in a study in progress (Bond & colleagues, in progress). This study is comparing two vocational approaches. One is the IPS model of supported employment and the other has been labeled the "diversified placement approach" (DPA), which is a particular agency's vocational approach that has evolved over a 30-year period of time, borrowing from many sources and not aspiring to be a specific model for other agencies. The design of the research study requires the investigators to show that both programs adhere to their respective program approaches and that neither resembles the other. Whereas IPS emphasizes rapid job search and individualized consumer choice of jobs, the DPA approach uses group placements and a pool of job opportunities, suggesting clear criteria for model differentiation. In other areas (e.g., long-term support, integration with mental health services), the two models agree. Thus to demonstrate both treatment integrity and treatment differentiation, both programs must be rated on both fidelity measures.

What are some ways to discover the critical ingredients of an inadequately-defined approach? In these circumstances, inductive methods may be useful. In an inductive approach, the critical dimensions are discovered, using interviews and brainstorming with staff, observation, content analyses of documents, and other similar means. We describe five such methods: the Delphi Technique, concept mapping, ethnography, critical incidents technique, and content analysis.

Delphi technique

A method developed by the Rand Corporation to improve organizational decision-making provides a unique way of identifying the dimensions of a psychiatric rehabilitation model. This method, the Delphi Technique, systematically combines the opinions of various individuals into a single consensus using group discussion. Although there are variations on the Delphi Technique, the inductive approach involves generation of items by the panel.

Drebing and colleagues are using the Delphi Technique in an attempt to identify the key ingredients of Compensated Work Therapy (CWT), which is a vocational program within the Veterans Administration (Drebing & Van Ormer, 1999). An expert panel consisting of three experts was identified and asked to individually generate a list of potential variables important to CWT. This list will be distributed to each participant, who will individually rank each item on the list based on its importance to CWT. Finally, the panel will convene, as a group, to review the rankings and to compile the final list of components.

By combining both a group and individual approach, the Delphi Technique may have the best of both worlds. The generation of ideas individually eliminates the potential of "group think"--a problem encountered in many focus groups--and thus maximizes the number of items generated. The subsequent review of the ideas by a focus group, however, increases the reliability of the dimensions selected.

Concept mapping

Another way to use focus groups to identify the dimensions of a model is a process called concept mapping (Trochim et al., 1994). Concept mapping uses focus groups to identify the important dimensions of a model, but takes one step further by using multi-dimensional scaling and factor analysis to make the method more quantitative. Shern, Trochim and LaComb (1995) used this method to identify the dimensions of the Choices program, a psychiatric rehabilitation program designed for the homeless that advocates aggressive

outreach, identification of goals, and the reintegration of individuals into the community. Two focus groups were established. The first group consisted of eleven case managers from the Choices program and the second consisted of nine Boston University experts in psychiatric rehabilitation. Each group's brainstorming lists were eventually constructed into a concept "map" through various factor analysis and scaling procedures. These maps were then combined to determine the components with the most agreement.

Concept mapping is suitable for discovering the underlying structure for a program approach that has not been formally conceptualized as a model. Thus, one advantage to concept mapping is that the brainstorming technique eliminates the need to establish preconceived notions about the key dimensions used in the structured interview approach. The potential disadvantages concern the subjective nature of focus groups, not only regarding the selection of experts, but also concerning group processes and the necessity to have adroit leaders to ensure that all voices are heard.

Ethnography

Ethnography is a method of discovery based on participant observation and open-ended interviews. Its appeal lies in the fact that observers see first hand the workings of the program model. Another appealing feature is that it involves obtaining information directly from program participants. Although promising for that reason, ethnography does have its drawbacks. Ethnographic discovery of critical program features may work well for well-established, successful programs that are following a well-defined model, but it may not be an appropriate method for new models or newly established programs. Ethnographic approaches can provide very different views of a program model than those intended by the program originators. The classic example of this difference is found in the ethnography of the original ACT program (Estroff, 1981).

Ware, Tugenberg, Dickey and McHorney (1999) used ethnographic methods to study the continuity of care in mental health services at two community mental health centers and one emergency psychiatric evaluation unit. Ware and her team personally visited the three sites where they observed team meetings and observed client contacts. In addition, they interviewed the supervisors of the program as well as the case managers. Through these observations and open-ended interviews structured around broad concepts rather than specific questions, they were able to identify six underlying themes of continuity of care.

Drawbacks include the fact that ethnographies are costly and time-consuming. More than structured methods, ethnographies yield unpredictable results and are highly dependent on the skill level of the ethnographer. Other assumptions that may or may not be appropriate include the assumption that the site or sites being observed are exemplars of the program model and assumptions about the "expert" status of respondents. Another consideration is that ethnographies reflect the opinions, perceptions, and assumptions of those who are observed and interviewed (as is true for other methods).

Critical incidents

A job analysis technique, normally used for identifying the critical requirements of a job, provides another possible approach to identifying the key dimensions of a program model. The idea behind the critical incident method is simply to identify behaviors that separate effective employees from non-effective employees. By analogy, successful programs in theory can be differentiated from unsuccessful ones by identifying what they do differently. Thus, this method can be adapted to the discovery of critical ingredients, by identifying those behaviors that separate effective programs from non-effective programs. Although an exact example does not

exist in the psychiatric rehabilitation field, critical incidents have been used in other psychiatric domains. Yalom (1985) used a critical incidents technique to determine the critical components of group psychotherapy. He asked his patients to describe the "single critical incident most helpful to them." Through careful reading, he then classified them into 11 primary factors: Installation of hope, Universality, Imparting of Information, Altruism, Corrective Recapitulation of the Primary Family Group, Development of Socializing Techniques, Imitative Behavior, Interpersonal Learning, Group Cohesiveness, Catharsis, and Existential Factors. Subsequently, he had patients rank order these in importance. His framework has proved to be useful in the study of a wide range of therapeutic groups.

This critical incidents technique is appealing, because of its emphasis on specific incidents, rather than broad attitudes. The greatest challenge in this approach is to sift through the critical incidents and intuit the program dimensions of interest.

Content analysis

An intriguing inductive method was employed by Wang, Macias and Jackson (1999). They reasoned that the results of sites from a formal accreditation process would yield important information about the critical ingredients differentiating well-implemented programs from those needing improvement. They reviewed the routine certification site-report visits of 15 New York City clubhouses. Using these reports, Wang's research team constructed content codes by identifying words, sentences, or phrases that appeared to capture a conceptually distinct evaluative criterion. Once the concepts were identified, each report was marked with a + or – denoting a positive or negative reference to the identified concept. The total score for each site was based on the presence of a reference to one of the identified concepts. The final list generated for the key components of the clubhouse model was based on the criteria most frequently used, and the criteria with high discriminatory potential.

The appeal of this method is that it is based on actual case examples, suggesting that it will generate dimensions of relevance in "the real world." The method, however, presupposes a clear set of criteria on which the site visits are conducted, which raises the question of whether these criteria themselves are the logical starting point for developing a fidelity measure. This methodology appears to highlight the most salient features that are problematic for programs seeking accreditation. This can be particularly beneficial for new programs hoping to avoid the same traps into which other programs have fallen. This method also assumes that a sample of written site visits is available and that the visits have been completed in a consistent, reliable and valid fashion. Most psychiatric rehabilitation models have not achieved this stage of evolution.

Summary

Identifying the key dimensions of a model is a critical step in the development of fidelity measures. However, there are as yet no accepted standards for identifying dimensions. The methods outlined above by no means compose a complete list of the options that may be available. In the interest of giving a simple template for other researchers to follow, we recommend adapting the McGrew and Bond (1995) methodology, with suggestions gleaned from the two related survey methods (Marty et al., under review; Schaedle & Epstein, 2000), as a guide that will be appropriate in many contexts. The methods used by McEvoy et al. (1999) also are well worth consulting. When faced with the task of documenting the critical ingredients of an

approach that is not well specified, there are a number of creative strategies than have been used or could be devised. To recapitulate, the main steps are:

- Start with summary papers on program principles and practice guidelines, if they exist, and consult developers of the model, if appropriate.
- Make site visits to programs exemplifying the model.
- Convert basic principles into a series of items.
- Consult experts and reviews of literature on the model.
- Conduct a pre-pilot of the expert survey to improve wording and length, determine interview flow, eliminate redundancy, etc.
- Determine which group or groups are the best informants for identifying the model dimensions.
- Develop a systematic sampling plan.
- Develop a methodology for obtaining expert opinions.
- Collect data.
- Analyze data and/or implement process to identify consensus items.

Step 4 Determine if Appropriate Fidelity Scales Already Exist

Should evaluators use an existing scale, modify an existing scale, or create a new scale? There are three issues involved: How closely are the evaluators seeking to replicate an existing program model, how well defined is that model, and how adequate are existing fidelity instruments.

Using an Existing Fidelity scale

If replication of a well-defined program model is the goal and if there is a satisfactory fidelity instrument, then the recommendation is to use that instrument. The Appendix contains a compilation of psychiatric rehabilitation fidelity and related measures currently in use or in development. Unfortunately, as suggested in the Appendix, none of the current instruments have fully satisfied all the criteria for a completely psychometrically adequate instrument! Therefore, if an adaptation of an existing instrument is chosen, it will have been only partially validated.

An example of an application in which one is likely to use or modify an existing instrument concerns the evaluation of a well-defined program model, such as skills training. For instance, if the program was seeking to replicate the Liberman skills training modules, and the intent is to replicate exactly as the originators had in mind, then it would be reasonable to use the existing fidelity scale (Wallace et al., 1992). This is what was done in one statewide survey (Bond et al., 1999b). A second example illustrates instrument adoption with minor modifications, useful when programs modify a small number of features of a model. Zahrt et al. (1999) used 25 of 26 items on the Dartmouth Assertive Community Treatment Scale (DACTS), dropping one item on substance abuse groups, because that was not a feature of the particular service model being evaluated.

Often evaluators cannot resist the temptation to improve an existing scale, even when the intent is to replicate previous work, that is, when the program model is the same as in a previous study. Our recommendation

is that the evaluator retain all of the original fidelity scale items for comparative purposes, even if new items are added.

No Current Fidelity Scale Exists, but a Related Fidelity Literature is Available

Sometimes there will not be a specific scale available that suits the program model of interest, but there may be a related literature. So, for example, if the psychiatric rehabilitation model of interest is provided in a counseling format, then the logical step would be to consult the vast literature on fidelity measurement in psychotherapy, as suggested in Chapter 1 (e.g., Moncher & Prinz, 1991; Waltz et al., 1993). If one is developing a fidelity scale for a case management approach, then the DACTS comes to mind as a starting point.

Where to Start if None of the Existing Literature Fits

If neither of these suggestions fit, then the evaluator is faced with generating a new scale. Chapter 4 discusses the next steps in this process.

Chapter 4. Scale Development

In developing a fidelity measure, we assume that many evaluators are interested in developing linear additive scales (Nunnally & Bernstein, 1994), which is the time-honored way of measuring constructs in mental health services, familiar to most researchers. We are referring, of course, to a measure consisting of a series of items scored on an ordered response, in which the items are summed to give a total score. Symptom measures, such as the Brief Psychiatric Rating Scale are an example of this measurement method. We also assume that individual items are often of interest as well, as are subscales, as we shall describe. In Chapter 4, we will discuss in depth the process of developing a fidelity scale including:

- Formulating a scale plan
- Choosing data sources
- Developing items
- Developing response scale points
- Determining item order
- Data collection
- Training interviewers

Step 5 Formulate Fidelity Scale Plan

When creating a scale, it is important to have a fidelity scale plan. A scale plan states the model dimensions (as discussed in Chapter 3), gives definitions, and outlines the number and possible content of items required to tap into those definitions. Creating the scale plan serves three purposes. First, the plan will assist in developing items that are consistent with the model. The plan is essentially the road map used to guide the development of the measure. Secondly, it increases the likelihood that all important aspects of the model will be assessed. In other words, having a plan will help identify holes or deficiencies in the scale or existing scales. For instance, suppose that during the piloting, the evaluator discovers that rural programs score lower on an item about access to public transportation. The item, which was created with an urban program in mind, unfairly penalizes rural programs in which mass transit is not available. Having identified this item as problematic, the scale plan can be reexamined to identify what the item was intended to measure.

Identification of a problem item may result in changing the item or it may result in modifying the scale plan. Finally, a plan can be useful as a check or as information for others who question the appropriateness or usefulness of the scale as a measure of program fidelity. This documentation will help to communicate to others how specific items correspond to the program model.

The scale plan starts with model dimensions, which were identified in the preceding phase of the research. In this example, the IPS model is conceptualized around three dimensions – Staffing, Organization, and Services. Then, within each of these dimensions, the evaluator must identify items. In this example, one item refers to caseload size. This example illustrates fidelity scale using a general dimensional framework. An alternative strategy would be to use program principles as the basis for the scale dimensions and then proceed to define items related to each principle. In other words, a fidelity scale for the IPS fidelity model could have been con-

structed around the principles outlined by the model developers (Becker & Drake, 1993), generating items for each principle. Of course, the a priori structure generated by the scale plan may not correspond to an empirically validated factor structure (e.g., Teague et al., 1998), but the scale plan gives a framework.

In developing a scale plan, the evaluator must decide on the number of dimensions to use and the number of items within each dimension. A rule of thumb is that the number of items on each dimension is the same, but this sometimes proves to be impractical when some dimensions are more delineated than others are. If this occurs, then this may be a cue for more explication of parts of the model. The total scale length requires a careful balancing act between completeness and practicality. A scale that is too lengthy poses difficulties in terms of feasibility, respondent attention span, and other factors. A scale that is too short may not be comprehensive or reliable.

An item has two elements: an item stem, which asks the question (e.g., what is your caseload size?) and the response scale, which consists of the set of response alternatives that each question is scored on. Although there are many different kinds of response scales, we recommend an ordered response scale, because such responses can be more easily converted to a linear additive scale (Nunnally & Bernstein, 1994). As way of orientation, Table 4.1 offers sample items from a 33-item measure called the Quality of Supported Employment Implementation Scale (Bond et al., 1999c). In this example, the dimensions are Vocational Staffing, Organization, and Services. The items are Agency focus, Screening policy, and Rapid search. The response scales consist of 5-point behaviorally-anchored scales.

Table 4.1. Sample Items from the Quality of Supported Employment Implementation Scale (Bond et al., 1999c)

Subscale	Item	Item Point Anchors				
		1	2	3	4	5
Vocational Staffing	VS1 - Agency focus on SE: Ratio of vocational staff solely devoted to SE to all staff devoted to vocational services (including prevocational employment, clubhouse, agency-run employment, etc.). If staff time is split, then estimate % of time.	< 25% of total staff devoted to SE	26-50%	51-70%	71-90%	91-100% of total staff devoted to SE
Organization	O7 - Screening policy: Program does not have exclusionary eligibility requirements relating to presumed job readiness, such as substance abuse, violent behavior, minimal intellectual functioning, mild symptoms, or treatment compliance.	Screening criteria have clear intent of excluding poorer functioning individuals		2 or more exclusion criteria, but intent is still to include most clients with SMI		Consumers are not screened out because they are viewed as "not ready" or "lower functioning"
Services	S5 - Rapid search for competitive employment: The search for competitive jobs occurs rapidly after program entry.	First job application is > 1 year	7 months - 1 year	3 - 6 months	1 - 2 months	First job application is typically < 1 month

Step 6 Develop Items

What Makes for a Good Fidelity Item?

Obviously, fidelity items should reflect the model principles. Some specific features that are important include the following:

- **Items should refer to the structure and activities of the program and behaviors of the staff.** This seemingly trivial statement is intended to distinguish fidelity items from items that do not measure program implementation, such as program outcomes. Thus, for example, we do not consider the item "At least 50% of consumers work in competitive employment" a fidelity item, because it refers to an outcome. However, we do think the item, "At least 90% of staff placement efforts are dedicated to finding competitive employment opportunities" is a fidelity item, because it explains how the staff defines its work.

- **Items should refer to things under the control of the program staff and program administration.** Although external factors influence the excellence of a program, we do not consider them as part of the measurement of fidelity. Thus, for example, we do not consider the item, "A low unemployment rate in the community" a fidelity item, even though it may affect a vocational program's success. We also do not consider the item, "Support and cooperation from higher administrative levels" an element in program fidelity, because it is outside the control of program staff Administrative support is part of the organizational climate. (We do believe administrative support is critical to the success of a program (Becker, Torrey, Toscano, Wyzik, & Fox, 1998a)). Finally, we do classify items referring to staff turnover and staff capacity as elements of program fidelity (Teague et al., 1998), although it is true that neither are fully under the control of the program. It is, however, difficult to implement a program model fully if the program does not have adequate qualified staff.

- **Items should be written to fit with the sociocultural context.** In generating items for a drop-in center measure, we initially included an item regarding proximity to public transportation, having urban centers in mind (Bond, Evans, & Resnick, 1998b). During our pilot, respondents from rural areas pointed out the fact that rural communities do not have the same network of public transportation. We therefore modified the item to fit the underlying principles of accessibility. In some instances the best solution may be to use alternate forms of specific fidelity items for different sized communities.

- **Items should be clear and specific.** As we discuss in Step 8, fidelity items are rated from a variety of sources, including interviews with program staff. Regardless of the data source, clearly worded items will improve the reliability of the data collected. Fowler (1995) provided five characteristics of good items that lead to the collection of meaningful, useful information:

  - When asked, does everyone understand the question in the same way and is that consistent with how the evaluator intended it to be interpreted?
  - Is the item administered consistently?
  - Is what constitutes a reasonable answer consistently communicated?
  - Can the respondents get or have the information needed to respond?
  - Are respondents willing to divulge information?

Writing a Good Item

Keeping in mind the five characteristics of a good item, once the evaluator has determined how many items will be needed for each dimension, the item stems for each item can be developed. Some tips for writing good items are shown in Table 4.2 and discussed below. These suggestions are worded as if the fidelity item is being obtained from an interview, although similar guidelines apply to the instructions one would give to an observer or to someone who is coding data from a client chart.

Table 4.2. Tips for Writing Good Interview Items

1. Ensure item does not have multiple meanings.
2. Use slang and jargon sparingly, only when necessary or helpful.
3. Avoid "double-barreled" questions.
4. Use neutral wording.
5. Specify response category (e.g., minutes, hours, days, etc.).
6. Interviewer should perform any required calculations after the interview.
7. Assist the respondent as much as possible in retrieving accurate information.
8. Use smaller time frames (e.g., 1 month vs. 1 year ago) when asking respondents to recall.
9. When appropriate, ask factual questions rather than asking for opinions.
10. Attempt to obtain first-hand information.
11. Clarify that respondent is familiar with terms and concepts of the interview.

1. **Is the item written so there is only one meaning?** One common pitfall is to use jargon terms that may have different interpretations. For example, terms such as natural supports, assertive outreach, community integration, empowerment, and recovery have different meanings. In other words, does every respondent have the same definition of each component of the items, even "familiar" and obvious terms (e.g., case manager, employment specialist, vocational services, agency)? If different agencies or different individuals within an agency have an idiosyncratic definition of any of the focus words in the item stem, then responses will vary. To avoid this problem, all basic terms should be defined and, when a survey instrument is administered orally, the interviewer should make sure that the respondent is using the terms in way intended. As much as possible, use the respondents' own terminology when it is identical to the terms intended in the fidelity measure. If respondents refer to clients as "members," then be sure to use substitute the word "member" in the rest of the interview protocol.

2. **The language idiosyncratic?** The evaluator needs to be careful about using slang or jargon that will lead to multiple definitions or misunderstandings. Respondents may be uncomfortable asking the evaluator to clarify a term and could respond based on their guess about the meaning of a question.

3. **Each question should be focused on a single issue**. In other words avoid "double-barreled" questions. For example, the question "Do case managers and employment specialists spend enough time with clients?" has two questions embedded in it. If a respondent answers yes or no to this item, it is not clear if the individual is responding in reference to both or just one of the questions.

4. **Items should be worded as neutrally as possible**. For instance, an item which begins, "Does your program follow the ACT guidelines for two contacts per week with each client?" is an exaggerated example of a leading question that may lead respondents to provide what they believe is the appropriate answer,

rather than what is really happening in their agency. The use of value laden terms, such as "consumer driven" may also trigger socially desirable responses. We were reminded of the demand characteristics of fidelity scales when we heard that one instrument had been dubbed the "The Correct Answer Is 5" Scale by clinicians completing the instrument.

5. **Questions should be worded in a way that respondents understand the form of answer requested**. For instance, if the question "How much time does a staff member spend on a particular activity" is administered, make sure that they know how the response should be in given (e.g., hours/per week). A better question would be, "How many hours does a staff member spend on this activity each week?" If the interview requests specific numerical information contained in records (e.g., client caseload size, etc.), respondents should be told this prior to the interview, so they can come prepared with client rosters and other relevant information.

6. **A fidelity item may require the determination of a percentage, rate, or other calculation**. The interview should minimize the need for the respondent to make calculations. For example, it may be simpler to ask how many hours a week a psychiatrist works on the program, rather than asking the percentage of a full-time equivalent position the psychiatrist is. Whenever possible, ask for the information that is necessary to make the calculation, rather than asking the respondent to do so.

7. **When asking for factual information, use strategies to increase accurate recall**. Respondents may be more motivated to be accurate if they fully understand how the information will be used. Also, accuracy will often increase if questions are phrased in particular ways. For instance, providing a time frame for the behavior (e.g., in the last 6 months, how many times…), asking the respondent to provide the behavior and the date it occurred, and providing the respondent with a list of events or behaviors to aid recall are all effective ways to increase recall and accuracy (see Fowler, 1995 for a complete review).

8. **Use smaller time frames to increase recall**. When asking respondents about the frequency of behaviors or events, they will have better recall if asked about shorter time frames (i.e., within the last month) compared to longer time intervals (i.e., within the last year) (Mueser et al., 1995).

9. **When appropriate, ask factual questions, rather than asking for opinions.** By asking questions that call for concrete information, the responses are more likely to be data-based and less likely to be colored by what the respondent thinks how the program should be functioning or what the respondent thinks is the "right" answer. Asking questions that focus on specific people, events, and activities also can lead to improved recall. In addition, the more specific a question, the less chance it has of being misinterpreted or misunderstood.

10. **Consider whether particular respondents have the information themselves or whether they will be answering based on second-hand knowledge.** It is best to make sure that the right people are asked the appropriate questions. Therefore, in the scale plan it might be useful to consider which items can be most accurately answered by each respondent. This may occur during the item review, when experts and staff members can help determine who can best respond to each item.

11. **It is important not to make assumptions about the knowledge respondents have that would be needed to answer the items.** For instance, consider the question, "With the changes in X policy, do you think….". The respondent may or may not be familiar with the policy change or the respondents may not

be familiar with the evaluator's terminology. Thus, their responses may not be interpretable. They may be uncomfortable asking the researcher to describe this change, they may be embarrassed to admit that they do not know the policy, or they may assume they understand the question and simply answer. Either way, the response received may or may not be appropriate to the information that was desired. (Do you want to add anything about the willingness of respondents to divulge potentially negative information?)

Although following these suggestions should increase reliability and validity, the true test will come in the pretest or pilot measure. Often times even a very carefully developed item will have problems that are not identified until the pilot phase.

Step 7 Develop Response Scale Points

What Makes for a Good Response Scale?

We recommend response scales with these attributes:

- standard number of scale points for every item (We recommend 5 scale points)
- ordinal scale points approximating equal intervals between each point
- behaviorally-anchored
- no gaps in the response alternatives
- no overlap in the scale points
- scale point based on the empirical literature

We recommend the use of behavioral anchors for response scales. In developing a new fidelity measure, evaluators should exercise great care in identifying anchors that match as much as possible the natural response pattern of the respondent. Ideally, the distribution of responses across programs should include all the scale points in the response scale. There is an art to anticipating the possible distribution of responses (as well as the level of precision at which distinctions can be made), and in the initial instrument development, the evaluator often must depend on guesswork, making refinements after piloting. Normative data are always a help in determining scale points. So, for example, in determining the scale points for client:staff ratio in a case management program, it is well-known that traditional case management programs may have case load ratios as high as 100 or more, whereas caseload sizes of 10 or less are typical of ACT programs. In this example, a 5-point scale is relatively easy to construct, but other examples are less obvious. Some other suggestions are as follows:

- **We recommend using ordinal scale points with 5 response alternatives.** Some researchers recommend up to 7 alternatives. Using more than 7 response alternatives is unnecessary for most rating tasks, on the assumption that respondents cannot make any finer distinctions. Moreover, if a scale has too many response options, respondents may find it difficult to keep the options in mind. Using fewer than 5 points may be cruder than is necessary. In summary, studies are needed to determine the optimal number of scale points.
- **Whenever possible, the number of response alternatives should remain consistent throughout a scale.** If a 5-point response scale is used, such as in the DACTS (Teague et al., 1998), then all items

should be scaled in that same format. However, we should note that some items are dichotomous (e.g., presence or absence of something), and do not lend themselves naturally to finer distinctions.

- **How to score dichotomous items.** In the case of a dichotomous item used as part of a scale comprised of items mostly scored on a 5-point response scale, the dichotomous item can be scored Absent = 1, Present = 5.
- **Anchor points should be meaningful and understandable to the respondent.** Again, if jargon or ambiguity defines these response options it will negatively influence the usefulness of the responses.
- **Response options should not overlap**. Similarly, there should be no gaps between alternatives (e.g., 10-20%, 25-35%). Early versions of the DACTS violated this principle, causing confusion in ratings (Winter & Calsyn, 2000).
- **In constructing the response alternatives, we recommend that the highest level of the response scale allow some leeway**. Thus, for example, instead of saying "100% of consumers are placed in competitive jobs," we recommend the wording, "90% or more..." This is in recognition of the fact that in the real world, rigid compliance is seldom achieved and in fact sometimes occasional exceptions to the rule are part of the model.
- **Some attention should be given to social desirability and response bias.** That is, the respondent should not conclude that the "best response" is always a particular response alternative. Some measurement experts suggest that scales be balanced with an equal number of positively worded and negatively worded items (Fiske, 1971).

In general, we encourage researchers to use empirical evidence to guide the development of behavioral anchors. This is consistent with Sechrest, West, Phillips, Redner, and Yeaton (1979), who have described the following three "appropriate standards of comparison" in program implementation (p. 531): 1) an "average" criterion based on normative conditions in other programs, 2) a criterion based on the identification of an "ideal" program as specified either by the authors of the approach or by participants and staff, and 3) theoretical analysis and expert judgment of goodness of fit. Which standard is best and under what circumstances, however, has yet to be established (Scott & Sechrest, 1989).

In addition to empirical evidence, it may be helpful to use model/program experts to develop the behavioral anchors. Model experts could be useful in identifying 1) the optimum response to each item to determine a "full and complete adherence to program model," 2) the possible range of responses, 3) behavioral anchors to maximize variability on the scale, 4) behavior or circumstance that best defines a particular dimension. Each of these is important to assure usefulness of a response scale (McGrew & Bond, 1995).

Using model experts can also help to obtain the range of possible responses. As noted above, the goal is to develop items and response scales that will accurately capture the variability in programs. If the behavioral anchors do not include all of the possible responses, or if only part of the range captures all of the possible responses, then the response scale is not appropriately sensitive. It is best if all responses have the potential of being chosen. For instance, using the team meeting example, if it is unlikely that teams never meet, than having "never meet" as the lowest behavioral anchor really is useless. If all teams meet at least once a week, than once a week rather than never is probably the more appropriate lower end of the response scale.

The high and low ends of a response scale are often the easiest to develop. It becomes more difficult, to identify independent but meaningful middle options. Because few programs are perfectly implemented or not adhering at all to program specifications, it is the middle alternatives that will be crucial for increasing the variability and thus the discriminability of the items. In addition, respondents/researchers will need to be able to differentiate between the options so that the differences will be meaningful to them. If the alternatives are too finely differentiated, respondents may have a difficult time choosing one. However, if they do not discriminate finely enough, all respondents could choose the same option, although their programs differ widely on the dimension being measured.

Step 8 Choose Data Collection Sources and Methods

Data collection refers to the information gathered in order to make fidelity ratings. Various data sources may be used, including staff, consumers, and charts. Different methods can be employed, such as interview, self-administered questionnaires, and direct observation. Some common data collection methods are given in Table 4.3.

What Makes for a Good Data Collection Strategy?

There is an art to devising a successful data collection strategy for a fidelity scale. Evaluators must consider practical issues related to each data collection method such as the time and resources required to access the source. For some methods, the resources spent to obtain the information may outweigh the usefulness of the information. For instance, chart reviews are labor intensive and do not always yield useful information, especially if charts are not monitored for accuracy.

Table 4.3. Common Data Collection Methods		
Methods	**Typical Advantages**	**Typical Disadvantages**
Interviews with Stakeholders		
Staff interviews (e.g., face to face or telephone)	Most efficient for simple information Allows for follow-up questions	Respondent biases Accuracy may depend on respondent's role Requires training of interviewers May be time-consuming
Consumer interviews	Actual recipients of services	Labor-intensive Problem of sampling Consumers not privy to "behind the scenes" activity
Collateral interviews	Unique perspective	Difficult to collect Biased, limited information Time intensive
Self-administered checklists (any of the above respondents)	Inexpensive Potentially quick turn-around	Variable completion rates Respondents may misunderstand questions, complete hurriedly or carelessly
Observation		
Site visits	First hand perceptions	Labor-intensive May be intrusive Staff and consumers may "put on show"
Observation and detailed coding of counseling sessions (on site or via videotapes)	Potentially highly objective	Labor intensive Requires training of observers May generate too much data Session(s) observed may not be typical
Job shadowing of clinical staff	Direct observation of actual staff behavior	Labor intensive Potentially intrusive Time sampling may not capture typical behavior
Record Review		
Manual review of clinical charts and agency records	Potentially objective data, congruent with agency "reality"	May not be complete, accurate, or up-to-date May not fit with evaluator goals May be difficult to access Client and staff confidentiality may be a barrier
Computerized management information systems	Data compilation may be simpler than manual chart review	Similar to manual review of records

Program administrators who are removed from the day-to-day activities of a program's operation often appear to have an idealized and therefore inaccurate view of how a program actually works, although they usually have the best information staffing patterns and start dates, for example. Chapter 5 gives empirical examples of agreement between sources. Some criteria for devising one's strategy are highlighted in Table 4.4.

Table 4.4 Tips for devising data collection strategy

- Choose data sources and methods that are congruent with the information needed
- Sources and methods may vary from item to item
- Use multiple sources and methods for each item whenever possible
- Train interviewers and observers
- Use multiple interviewers and observers
- Build in methods to check data quality
- Assess fidelity at more than one time point
- Assess reliability of ratings

Some examples of the possible choices in data collection procedures for fidelity scale in use are given in Table 4.5. These examples are intended to illustrate logical data sources and methods of data collection for specific types of information. Some fidelity items are strictly factual and easily accessible, such as the number of team meetings held every week. In this case, the best source would be some form of permanent record, such as calendar used to record staff activity. Even in this simple example, however, some intermediate steps are necessary, such as defining the time frame used to determine a program's "typical" frequency of meeting. Implicit in Table 4.5 is the congruence between type of program model and mode of data collection. Thus, for example, skills training involves specific therapist interventions within the context of counseling sessions. Accordingly, fidelity measurement for this type of program model reasonably involves direct observation. For careful fidelity measurement in this area, we would look to the extensive psychotherapy fidelity measurement literature for ideas. Conversely, for an item relating to professional distance between staff and consumers, one might use direct observation and interviews with consumers as the data sources.

The ideal data collection model includes the use of multiple data collection methods for each dimension. If all of the methods have high agreement, this type of triangulation can provide strong evidence for the validity of the fidelity measure and provide confidence in the data collected. A common outcome, however, is that sources and methods do not agree completely (Winter & Calsyn, 2000; Zahrt et al., 1999). In this case, the evaluator must decide which source (if any!) is the most credible. Another strategy may be to discuss the discrepancies among the available data sources (e.g., case managers and team leaders). These sources may be able to help the evaluator understand the discrepancy or reconcile it.

Another reason for using multiple perspectives is that they can inform each other. For instance, visiting a program and observing program procedures along with reviewing charts prior to conducting interviews may be maximally useful. If there are observations or chart notes that require clarification, it could be requested during the interview. Additionally, having those observations may also help to better code the respondents' answers to the interview questions and help to ask better questions.

It would be helpful to have evidence to guide the choice of data collection methods. Unfortunately, there appears to be little empirical evidence suggesting that which type of data collection method is most reliable and valid for these purposes.

Interviews

If interviews are used, the evaluator must decide who to interview and how to conduct the interviews. Possible respondents include: team leaders, staff members, program directors, state-level program monitors, researchers, outside site-visitors, and consumers. A basic principle is to obtain information first-hand from individuals who are most knowledgeable. In terms of interview styles, there are three basic types of interviews: structured, unstructured, and semi-structured.

With unstructured interviews, the interviewer may have specific information to gather, but either allows the respondent to direct the interview or uses unique follow-up questions in each interview. A large body of research suggests that unstructured interviews have low validity for decision-making (e.g., personnel selection) (Conway, Jako, & Goodman, 1995; McDaniel, Whetzel, Shcmidt, & Maurer, 1994). The drawback to unstructured interviews is that information is not consistently obtained and, therefore, the picture one obtains of a program will depend on the direction an interview takes. A structured interview refers to a very specific and rigid interview schedule in which the intent is to standardize the process, including the ways: 1) the interviewer introduces the task, 2) questions are posed to the respondent, 3) questions are probed in a follow-up, 4) answers are coded, and 5) the interactions are structured between the interviewer and the respondent (Fowler, 1995). A semi-structured interview refers to an interview that has a moderate to low level of structure, but also allows the interviewers to follow-up with questions and information that arise during the session. Rather than placing a rigid structure on the process, as in a structured interview, the interview is guided by a set of pre-determined goals for information collection. For the purposes of rating fidelity scales, structured or semi-structured interview formats are more appropriate than unstructured interviews.

Table 4.5. Guide to Data Collection Strategies for Illustrative Fidelity Items

Example	Type of Model Where Used	Inference/Ease of Obtaining Accurate Data	Types of Item	Intermediate Steps	Suggested Sources for Data
Number of team meetings per week	Case management/ voc programs	Low inference, Easy to obtain	Strictly factual, no calculation	Define time frame (e.g., last 3 months).	Written record of staff meetings
Caseload ratio	Many models	Low inference, Easy	Strictly factual, some calculation	Define time frame and caseload. Identify direct service staff	Team leader report of number of full-time equivalent direct service staff; Roster of clients
Specific program service, such as prevocational work crews	Vocational	Low inference, Easy	Presence/absence of specific program component	Define terms concretely, provide alternative labels (e.g., "work units," "work readiness training").	Ask program staff
Consumer choice primary determinant of job selection	Vocational	Moderate inference, Moderately hard	Interpretation of service provision style	Define what is meant by individualized job search and by typical alternatives.	Ask program staff to list all working clients on caseload. For each, explain how client obtained job. Continue probing until sure.
Professional distance of staff from consumers	Drop-in centers, clubhouse	Moderate inference, Relatively easy	Generalization of behavior pattern	Identify indicators (e.g., staff have nametags, carry beepers, have own offices, eat lunch away from consumers).	Researcher observation on site
Role-playing integral part of program	Skills training	Low inference, Moderately hard	Therapist behavior during counseling sessions	Define role-playing.	Researcher counts number of role-plays during a counseling session
% consumers working in agency-run business	Vocational	Low inference, Moderately hard	Summary of consumer activity	Define jobs that qualify. Define time frame	Management information system report
% of clients receiving two or more contacts per week	Many models	Low inference, Moderately hard	Frequency of service provider contact	Define clients included in sample. Define time frame.	Management information system report

Site visits

We recommend that site visits include two or more observers, to provide different perspectives and to permit examination of reliability of ratings. In addition, with at least two observers, it is less likely that important information will be overlooked or forgotten. As with interviews, we recommend structured or semi-structured protocols for rating fidelity, including the use of a site visitor inventory (Hargreaves et al., 1998). Examples of data that can be collected through this type of inventory include: space availability, staff facilities, and safety features. One advantage to having researchers code this information, rather than gathering it through interviews, is that researchers have been trained to view programs from the viewpoint of assessors rather than as staff members in the program. Additionally, having observers in a program could influence consumer, as well as, staff behavior. Researchers need to be careful to be as unobtrusive as possible to encourage typical behavior.

Charts/records

Institutional records and charts can be valuable for either a primary data source or as a way to verify other data sources. Records and charts might include client charts, daily logs kept by program staff, records from staff meetings, and institutional reviews. In addition, many organizations maintain computerized records in a management information system (MIS), which may include pertinent information such as diagnosis, service contacts, vocational status, and hospitalization usage. Clinician activity logs, documenting staff and program activities, have been used in several fidelity studies (Brekke & Test, 1992; Brekke & Wolkon, 1988; Teague et al., 1995).

Without exception, clinicians maintain client charts. Traditionally, clinicians often had little incentive for keeping accurate and complete records. Unless clinicians are trained to record in a consistent manner, the data are likely to be fraught with inaccuracies and missing data. Reviewing client charts is a time-intensive process, and may require an elaborate coding system in order to quantify chart data. In addition, client-level data is often not the level of data that is collected for the measurement of program implementation. Like charts that are maintained manually, MIS records vary widely in completeness, accuracy, easy of access, and timeliness. However, in principle, MIS records offer distinct advantages in terms of completeness of records and ease of access, compared to manual access to records.

Thus, the type of chart/record to be used for data collection will depend on the type of program being studied and what kind of information is needed to complete the measures. Charts completed by clinical staff will vary in reliability and validity as a function of their uses and incentive systems (Clark et al., 1994; Wolff & Helminiak, 1996). If certain kinds of services are billable, but not others, then the billable services will be more frequently recorded. Suggestions for improving charts as a data source include: regular internal audits of charts, training staff on charting to increase consistency, or creating highly structured charting forms (Clark et al., 1994). In addition, comprehensive coding systems can also help to ensure that the researchers obtaining information from charts and records are doing so in a consistent fashion and are gleaning as much data as possible from the source. Creating a new system for keeping chart information, such as implementing a daily staff log, may also be a way to ensure more valid and complete information.

Self-administered checklist/questionnaire

Rather than interviewing respondents, a self-administered checklist may be developed to assess fidelity. Advantages are that it is easy to administer, a fairly large number of questions can be asked, and respondents can respond at their own convenience. However, if items are ambiguous or unclear, the respondent will not have the opportunity to ask for clarification. When using a self-administered scale, response rates are typically low unless follow-up activities are used (e.g., reminder letter, second copy of scale). Using such follow-up activities will add to the length of the data collection process (Fowler, 1988).

In deciding the data source and method, there are more broad issues to consider. As a practical matter, if a questionnaire or interview is too long, source respondents may become inattentive, give less thoughtful answers, or fall into a response set pattern. Among the many contributing factors in determining the optimal length for an instrument are mode of administration, type of respondent, complexity of the questions, and the difficulty of the format. All things considered, an interviewer-administered questionnaire is likely to be more accurate than a self-administered questionnaire if the interviewers are well trained. Type of respondent may also influence the quality of the data. Program staff workers are often busy, and therefore may resist long instruments. On the other hand, program staff typically have a minimum level of education as a requirement for being hired, and reading level is usually of less concern. In addition, program staff have other paperwork requirements, making fidelity questions less unfamiliar. Some mental health consumers are highly motivated to help in completing surveys, but missing data are a common problem with this group. For example, Burt, Duke, and Hargreaves (1998) in piloting a 129-item questionnaire, reported a 48% rate of missing data. The complexity of questions is another factor influencing the response burden. Obviously, items requesting simple ratings are less demanding than questions requiring more detailed information.

Our experience in piloting various instruments suggests that instruments that are longer become increasingly taxing on both the interviewers and the interviewee. For self-administered surveys, we suggest that the survey take no longer than 30-40 minutes to complete. In addition, we suggest that evaluators realistically estimate how long the survey/interview will take, so that respondents can anticipate how long they will need. If the length of the survey/interview is severely underestimated, some respondents are likely to react accordingly and this may result in less credible data.

Step 9 Determine Item Order

The order in which items are placed on the measure is important. This is especially true if the measure is tapping sensitive information or if respondents are nervous or reticent about responding. In addition, sometimes the ordering can help a respondent's recall. Below are some suggestions to consider when designing the layout of the scale.

- **To ease the respondent into the process, ask innocuous, easy items at the beginning**. This will serve to relax the respondent, reduce any threat they perceive from the process and allow them to respond with the necessary information. Some suggest that more personal information (e.g., hospitalization history, income) should be placed at the end of the scale to reduce response apprehension. They believe that asking for this information at the very beginning of the scale may influence subsequent responses. However, if the interview has gone on too long, leaving important questions to the end may not result is accurate or useful responses.

- **Design the questions in a logical order.** One such ordering capitalizes on the fact that some programs involve a typical sequence of activities. For example, the sequence in some vocational program is: Intake, Assessment, Job Selection, Placement, and Follow-Along. Asking questions in this sequence will lend coherence to the interview.
- **Similarly, questions that are related (i.e., concern the same dimension or topic area) should be kept together.** Respondents may be confused or frustrated if the scale/interview jumps from topic to topic. For instance, if items about job placement appear at multiple points in the scale, respondents may interpret the latter questions as referring to a different issue. Respondents expect similar items to be grouped together and this will improve recall. A second reason for keeping items together is that if similar resources are used to answer the question (e.g., attendance charts, employment records, client records) it will be easier for the individuals to respond and will thus reduce their frustration.
- **Begin with general questions.** If there are multiple questions about a particular dimension or area, begin with more general questions and move to more specific ones.

Step 10 Develop Data Collection Protocol

Another way to enhance the standardization of the data collection process is through the development of a protocol. These protocols should include, in written format, all instructions and directions for the evaluator. Although protocols are typically developed for interviewing purposes, we believe that protocols may be useful for data collection through site visits, and chart reviews, as well as interviews. The purpose of such a set of instructions is to increase inter-rater agreement and standardization of data collection. Although the general purpose of the protocol is similar across data sources there are a few specific guidelines for interview protocols that we would like to provide.

For instance, the protocol for the interview should include the introductory information provided to the respondent, as well as any standardized phrasing for probing questions. If the interviewer is responsible for coding the responses, the protocol should also include an expanded, clearly written definition of each of the response scale anchors. Even an experienced interviewer may find it difficult to remember all of the decision rules for coding responses. In addition, a protocol may be reassuring to new interviewers and may lower the chance that procedures or items will be used in an idiosyncratic way. For instance, having introductory statements and items written out will lower the chance that interviewers will paraphrase or skip critical information. In summary, a well-crafted interview protocol may greatly reduce between interviewer variance (Fowler, 1988).

A protocol is also helpful to standardize data collection through chart reviews. As discussed earlier, charts can contain a wealth of information, but that information can be difficult to quantify. Clear guidelines should be developed to identify coding decisions and locations to find the relevant information. The protocol should also outline the number of charts needed and how they will be selected (e.g., random vs. convenience). As with interviews, protocols for chart reviews can increase the reliability of the information gathered.

Step 11 Train Interviewers/Raters

Attention to the selection, training, and monitoring of interviewers is critical to fidelity measurement. General principles for tackling these issues are widely available (e.g., Stouthamer-Loeber & van Kammen,

1995). Sometimes, the researcher and the rater will be the same individual or set of individuals. However, other times, raters will be selected from a pool of candidates. Raters who are independent of the development process are presumed to provide less b`iased ratings, because they may be less invested in obtaining findings consistent with preconceived notions of what is the best way to provide services.

Selection of interviewers includes an assessment of interpersonal and listening skills, objectivity, and critical thinking. In addition, an interviewer should have some background in the mental health field, knowledge about severe mental illness, and some working knowledge and exposure to the program model in question (Bond et al., 1997a). With this type of background information, the interviewers are better able to code responses and follow-up on incomplete or ambiguous responses. Interviewers should receive systematic training on general interviewing procedures and on specific features of the scale in question. This training should include instruction on the following (Fowler, 1995):

- How to contact and introduce scale to respondents
- The questionnaire layout and how the interviewer should progress through it (e.g., skip items)
- How to probe or follow-up if initial responses are not on track
- How to code responses and place them on the response scale
- How to interact with the respondent.

Other issues covered in the training may include the overall purpose of the scale and/or specific items, and how to handle issues like confidentiality or other respondent concerns. We recommend lectures and in-depth discussion about the scale, followed by thorough field training. An experienced interviewer should observe the field training and provide guidance and feedback to the trainees. Some tips for interviewers are given in Table 4.6.

Table 4.6. Tips for Interviewers

- Prior to the interview, inform the respondent of information that would be useful for the interview such as attendance rosters, charts, lists of employed clients, etc.
- It is important that the interviewer be familiar with the site/sample and the program. Become familiar with the language and jargon used prior to the interview.
- To make the process more productive, it is important to orient the respondent to the purpose of the interview at the beginning of the interview.
- Make accurate estimates about the amount of time required for the interview. It is better to overestimate the length than to underestimate it.
- To keep the interview on focus, it is important (if there are multiple interviewers) to elect an interview leader to provide direction.
- Make notes about extraneous information or observations. This information could be helpful when later trying to clarify responses.
- Encourage respondents to ask questions if they need clarification.
- Encourage respondents to contact the interviewer following the interview if they have identified new information they believe is relevant.

Chapter 5. Piloting the Scale

In this Chapter, we will describe how to test the newly developed scale in order to identify problem areas and assess reliability and validity.

Step 12 Pilot the Scale

The pilot implementation of the fidelity scale is critical in identifying problems with the sequence of the items, ambiguity in items, confusion in response options, or any other issue related to the scale or the data collection methods identified in its development. For instance, if data is collected through an interview, piloting also provides actual practice for the interviewer in asking the questions as well as coding the responses. We suggest that there should be two phases to the pilot scale. The first would focus solely on the content of the items and the second on the psychometric properties of the scale. This is consistent with Fowler (1995), who suggested that intensive interviews take place prior to the actual pilot. This allows researchers to assess how well respondents understand the items and what they are thinking about as they answer the questions. A good example of this process is contained in Burt et al. (1998), who used multiple pretests to develop a measure of consumer perceptions of community-based programs. We should acknowledge, however, that it is not always feasible or reasonable to conduct a two-step pilot process. When this is the case, we encourage researchers to seek out both content and psychometric information in the same stage.

Content Pilot

For the first pilot, we encourage researchers to find a small sample with which to pilot the items. The goals of this pilot are to (1) determine feasibility of data collection methods, (2) give interviewers, observers, and chart reviewers practice, (3) identify problems with the pace or placement of the items, (4) identify terminology or jargon problems, (5) identify whether the response scale is appropriate, and (6) assess whether the respondent has other information that would be vital.

Number and type of respondents for content pilot

Although circumstances, time constraints, and resources may suggest otherwise, we recommend as a rule of thumb that about 10 pilot interviews be conducted at this stage, using respondents from different programs. They should be as similar to the actual respondents as possible (e.g., job classification, similar program). However, this sampling strategy will need to be modified if the fidelity study is being conducted on a relatively unique program that has not been widely disseminated. In this situation, some of the same cases in both the pilot and actual implementation may need to be used. In sum, the goal is to pilot the items on a sample as representative and as large as possible given the practical constraints.

Type of information sought in the pilot

The pilot gives an opportunity to determine the feasibility of locating different types of information. If interviews are part of the fidelity measure, interviewers can use the pilot to obtain respondent reactions. One method is to review each item with the respondent to determine whether any problems existed with the item stem or the response scale. In addition, respondents can be asked whether there are other critical components to the program that were not assessed. This might also be a good time to find out if respondents believe they could adequately answer each question, or if another person on the team would have been a more appropriate respondent. These types of questions can be valuable in discovering if the measure is deficient.

After completing each of the pilot interviews, the evaluator should ask the interviewers about the interview format. First, the researchers should determine whether the interviewers felt comfortable with the process and answer any questions they might have. Secondly, the evaluator should determine if any respondents were uncomfortable giving sensitive information or if they struggled to answer certain questions. The interviewers' comments, along with those of the pilot respondents, can provide very valuable information for making appropriate modifications. This is best done immediately after the data is collected for each site, so that the information is fresh. We also recommend that the primary investigators be involved in the data collection during the pilot, rather than completely delegating this work to others. Finally, we encourage the use of two or more data collectors/interviewers during the pilot phase, to provide another feedback loop.

The researcher should also review the actual responses individuals gave and the coding done by the interviewer (if appropriate). If multiple programs are being used in the pilot, this review should also reveal whether there is good variability on the items. In addition, the researcher should determine whether the interviewers believed the ratings given to the various programs seemed to accurately discriminate between them. In other words, were response options able to capture the differences between programs that interviewers encountered?

Procedure for question revision

If items or response scales are identified that appear to cause problems for the respondents or the interviewers, modification may be necessary. Before changes are made, however, it is important to consider whether the problem was idiosyncratic to the respondent/interviewer or whether there is actually a problem with the item. If 10 pilot interviews are conducted and only one individual has a problem with a particular item, it may be too hasty to make changes. However, if a pattern emerges with several individuals struggling with the same item, modification should be taken more seriously (Fowler, 1988).

If an item is selected for modification, it is important to first review what the item was intended to measure. This will require that the researcher revisit the scale plan and examine the dimension and sub-dimension this item was meant to tap. After hearing the difficulties respondents have with an item, it may become clear to see how the item missed the dimension. Other times it may not be as clear and the help of an expert or perhaps one of the pilot respondents should be sought. For instance, the expert or respondent may be able to help identify how the item got off track if they understand the original purpose of the item (i.e., what it was intended to measure).

Psychometric Pilot

Sampling sites

A particularly thorny question in developing fidelity scales is determining the minimum number of sites necessary for psychometric analyses. If the intent is to conduct a factor analysis, some authorities recommend a minimum sample of at least 150 or 200 (Hinkin, 1995; Nunnally & Bernstein, 1994). For assessing reliability, investigators commonly use considerably smaller samples. As a practical matter, most fidelity studies have been conducted on considerably smaller samples. One factor analytic study used 50 sites (Teague et al., 1998), while a statewide survey included 76 sites (Bond et al., 1999b). Another factor analytic study, involving a cumulative data base compiled over several years, included 123 sites (Bond, 1999). However, many other psychometric studies of fidelity measures have used samples in the range of 18-32 (Bond et al., 1997a; Bond,

Picone, & Mauer, 1998c; Johnsen et al., 1999; Lucca, 2000; Winter & Calsyn, 2000) or fewer (Teague et al., 1995; Zahrt et al., 1999). The number of available site will significantly influence the potential sample size. If a study involves a statewide initiative (Teague et al., 1995) or multi-site study (Winter & Calsyn, 2000), then the number of sites of interest typically are limited to the number of sites participating.

The underlying principle is that the larger the sample, the more stable the statistical estimate. If the sample size is relatively small, the statistics will be less stable and likely to change with another or larger sample. Therefore, we caution researchers about making extreme modifications in their scale or drawing strong conclusions when their sample is small.

Other factors besides sample size affect the stability of psychometric analyses. One factor concerns the distribution of responses. Samples that are highly homogeneous (i.e., when there is a restriction of range problem) will have unstable reliability coefficients and unstable factor analytic structures (Bond et al., 1998c; Vogler, 1998). Thus, in a pilot it is useful to include some sites that are not exemplars of the program model of interest. Another related consideration in the adequacy of factor analysis and other analyses of internal scale structure concerns the mean size of the inter-item correlations (Guadagnoli & Velicer, 1988).

Type of information sought in the psychometric pilot

During the psychometric pilot it is still important to seek information about the clarity and appropriateness of items. In addition, we encourage researchers to investigate the length of the interview/data collection, look for redundancy in items and unnecessary items. Finally, it is at this point that the psychometric properties of the scale should be quantified. These analyses should be conducted both during the pilot phase and following the actual implementation.

Step 13 Assess the Psychometric Properties

The two properties that we are most interested in examining are the reliability and validity of the scale. Briefly, reliability refers to the consistency of the responses for a particular item or scale and validity refers to the degree to which the scale measures what it was intended to measure.

Table 5.1 provides a quick reference for distinguishing the different types of reliability and validity, however each will be described in more detail below. The following discussion assumes an elementary level of understanding of the topics and relevant issues related to the assessment of the psychometric properties of scales. We encourage readers new to this area to seek a basic statistical reference to gain introductory information.

Table 5.1. Quick Reference for Reliability and Validity			
Types of Reliability: Consistency of a measure		**Types of Validity: If scale is measuring what is intended**	
Internal Consistency	Average correlation among items	Face Validity	Items appear to measure what they are intended to
Test-Retest	Correlation between scores on two separate occasions	Content Validity	Items in the scale are in accordance with the model
Inter-rater Agreement	Correlation between two raters rating same program	Construct Validity	The scale is measuring the identified concept or program accurately
		Convergent and Divergent Validity	Fidelity scores correlate with other indicators/measures of the model of focus (e.g., certification, self-labeling) and are uncorrelated with indicators/measures of other models.
		Criterion oriented Predictive Validity	Programs that score high on fidelity measures have better outcomes in identified domains.

Item Analysis

The study of psychometric properties of a fidelity scale begins with an examination of item distributions. An example is shown in Table 5.2. In this example, caseload size shows little variability within the sample. In other words, almost all of the programs had low caseload ratios, as defined by the scale. By contrast, the item, "Contact with the mental health team," showed wide dispersion. The dispersion of an item is one criterion for its utility. Another common feature in item analyses is examination of correlations between the item and total scale and subscales (not shown in this example).

Table 5.2. Portion of Item Analysis for IPS Fidelity Scale Items (adapted from Bond et al., 1997)

			Interrater Reliability (ICC) (n=22)	Distribution of Ratings		
Subscale	**Item Label**	**Item Descriptor**		**Low (1-3)**	**Medium (4)**	**High (5)**
Staffing	Caseload size	Employment specialists manage caseloads of up to 25 clients	.99	11%	11%	78%
	Exclusively vocational	Employment specialists provide only vocational services	.95	19%	26%	56%
	Generalist model	Each employment specialist carries out all phases of vocational services	.93	22%	19%	59%
Organization	Contact with mental health (MH) team	Employment specialists are part of mental health treatment teams; routinely share decision-making with this team	.86	59%	11%	30%
	Vocational unit	Employment specialists work as a unit -- have group supervision and shared caseloads	.67	26%	33%	41%
	Zero exclusion	No eligibility requirements (such as job readiness) for program	.83	52%	30%	19%

Reliability

Reliability refers to the consistency of the responses for a particular item across time or consistency in responses to items within a given scale. If the responses received for any given item varied greatly across administrations it would be difficult to determine which was the most appropriate response. In other words, we would like to assume that any variability in responses is due to true differences in the program being measured, rather than time of administration or a particular quality of the item. Therefore, it is important to assess the reliability of measures.

Often times a rule of thumb has been used to determine the "minimum threshold" acceptable (.70). This threshold appears to have developed with the suggestions of Nunnally (1994). However, many factors influence the value for a reliability coefficient and, rather than simply seeking to reach that value, researchers carefully consider the type of reliability, the content of the scale, the respondents, the history of the scale, and the level of precision needed. Some scales (e.g., job satisfaction) routinely obtain internal consistency reliability coefficient values of.90 and higher. In this case, obtaining reliability of.70 would be a source of concern, even though it passed the.70 threshold. Conversely, if a scale is measuring a fairly new construct and the literature is not clearly articulated, reaching a value of.75 or lower might be a reasonable value for the initial use. Other factors like sample size, number of items, and the number of factors can also influence the size of the reliability coefficient. These issues will be more fully discussed below.

Methods of assessing reliability

There are several methods for assessing the reliability of a measure. When using fidelity measures, researchers have most often relied upon internal consistency, test-retest reliability, and inter-rater agreement as the primary methods. Each method has a different assumption about where error comes from. Thus, there is not one type of reliability (Pedhazur & Schmelkin, 1991). Rather, the choice of method will depend upon the type of error the researcher is hoping to better understand. A comprehensive approach to assessing different sources of unreliability is through generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). Generalizability theory uses an analysis of variance approach to differentiate between the various sources of variability (i.e., data sources, observers, and time). An example of the use of generalizability theory with a fidelity scale is given by Winter and Calsyn (2000). Generalizability theory is appropriate when different sources of variability have been assessed systematically (Nunnally & Bernstein, 1994).

Sources of reliability

We consider the following sources of reliability: (1) internal consistency, (2) test-retest reliability, (3) inter-rater agreement, and (4) agreement between data sources.

Internal consistency.

Internal consistency involves assessing whether all the items on a scale measure the same thing. The statistic most commonly used is Cronbach's alpha, which is based on the average correlation among the items (Nunnally & Bernstein, 1994). It also should be noted, however, that the number of items also heavily influences internal consistency. In using the criterion of internal consistency to assess the reliability of a fidelity scale, the evaluator assumes that the criteria, as measured by the fidelity items, "go together" as a group, so that well-implemented programs score high on most or all of the fidelity items, whereas poorly implemented programs or programs that are not following the model, do not. In our experience, we have found that internal

consistency is a useful criterion when one has sampled a wide range of programs (Bond et al., 1997a; Lucca, 2000; Teague et al., 1998), but it is a frustrating criterion when the sample is homogeneous (Bond et al., 1998c; Zahrt et al., 1999)

As part of the process of determining whether a scale is psychometrically adequate, the evaluator typically examines the contribution of each item to the internal consistency. Most statistical packages can report the overall coefficient as well as the coefficient if each item was deleted. Item inclusion is based on both theoretical grounds (Is this item so central to the model that it should be included, regardless of the psychometric consequences?) as well as psychometric grounds. Thus, the evaluator asks whether each item is jeopardizing the internal consistency. Small changes in alpha are not grounds for dropping an item; however, if dropping an item would lead to large gains in reliability (.09) the item should be closely examined to determine whether it should be dropped or modified. Similarly, items that are negatively correlated with the total scale score or with other items are problematic. It may be that they are negatively worded and should be reverse coded prior to the analyses, or they may just be poor items. In sum, researchers should be wary of items that are lowly or negatively correlated with the total score, as they will have a negative effect on the alpha coefficient.

Another potential influence on internal consistency is the stage of program implementation. Winter and Calsyn (2000) found that Cronbach's alpha for the DACTS increased from.57 to.72 to.82 over a 3-year period

in 18 newly-formed outreach programs. We speculate that low internal consistency may be more problematic for programs in their start-up phase.

Developers of fidelity instruments often conceptualize their measures as being multidimensional. Two general strategies for identifying internally consistent subscales are a priori methods, in which one postulates specific sets of items that are believed to hang together on theoretical grounds, and inductive methods, of which factor analysis is the most popular. In our experience, the a priori method of scale construction for fidelity scales has been mixed. An example of a successful factor analytic approach is given by Teague (1998), as shown in Table 5.3.

Table 5.3. Factors, percent of total variance, and principal items loadings for the DACTS (Teague et al. 1998)

Factor / Item	Variance / Loading	Factor / Item	Variance / Loading
Team & Intensity	14.2%	**Substance Abuse Treatment**	10.5%
Contact frequency	.82	Substance abuse groups	.83
Team meeting	.68	Substance abuse treatment	.78
Team approach	.64	Dual disorder model	.72
Service intensity	.64		
Small caseload	.63	**Specialist Staffing**	9.0%
		Substance abuse specialist	.78
Community Treatment	13.5%	Nurse on staff	.74
In-vivo services	.68	Psychiatrist on staff	.58
Hosp. admission responsibility	.63		
Treatment responsibility	.62	**Caseload Distribution**	7.4%
Admission criteria	.61	Team leader role	.84
Crisis services	.61	Intake rate	.63
Work with supports	.58		
		Staff Capacity & Continuity	7.1%
Engagement / Retention	10.8%	Staff capacity	.90
No-drop policy	.79	Continuity of staffing	.72
Assertive engagement	.73		
Hospital discharge planning	.69	**Vocational Specialist**	5.7%
Time-unlimited service	.57	Vocational specialist	.87

Having discussed this method of assessing reliability, we caution readers that we are uncertain whether internal consistency is truly the most appropriate method for obtaining evidence of the reliability of the measure. Fidelity measures are typically attempting to assess many varied aspects of a program. Thus, assuming that these items will be internally consistent, even within a defined subscale, may be an unwarranted assumption.

Test-retest reliability. The test-retest reliability coefficient is determined by the correlation between the scores of two separate administrations of the scale. The nature of the construct will determine whether this is an appropriate estimate of scale reliability. For instance, if the researchers believe that a program will naturally improve over time, then changes in the true score will occur. In a test-retest coefficient, these changes will be reflected as error. If no real changes in the program are assumed to take place over time, then test-retest is a reasonable option. Even if one expects over time, the repeated assessment of fidelity provides useful information about the scale's properties.

Another concern is carryover effects (Nunnally & Bernstein, 1994). Specifically, if the same set of items is given at two points in time, individuals' responses could be based on their knowledge of earlier responses rather than their accurate response at that particular time. Although no empirically defined rule has been identified, some researchers have agreed that a two-week period is appropriate for assessing test-retest reliability (Nunnally & Bernstein, 1994). We should note that in the mental health services research area, the test-retest time interval is often longer. Moreover, Nunnally and Bernstein also stated that the time period will depend upon each individual situation. For instance, in fidelity measurement, it is unlikely that programs will change over a two-week period, so having a longer period between assessments might be appropriate. However, a longer time period is chosen, it is more likely that intervening variables or true change will occur and negatively influence the correlation.

In the context of program fidelity, a test-retest reliability coefficient is typically labeled stability, to reflect the belief that change does not necessarily mean that the measure is unreliable, but rather that the program has changed. In the most extensive longitudinal study of a fidelity measure, Winter and Calsyn (2000) studied a group of 18 newly-formed outreach programs for homeless people with mental illness over a 3-year period. The stability of the DACTS was only .28. The authors noted that the low stability was a result of some sites changing more over time than others. With the long time interval the lack of stability is not surprising.

Inter-rater agreement. If multiple raters or observers are used to rate fidelity, then agreement between sources can be assessed (Pedhazur & Schmelkin, 1991). The two most common methods are the kappa coefficient and the intraclass correlation coefficient. The kappa statistic is appropriate when the ratings are categorical (Bartko & Carpenter, 1976). The intraclass correlation is appropriate when the ratings are continuous (Shrout & Fleiss, 1979). Table 5.2 illustrates the use of the intraclass coefficient at the item level. In that same study, intraclass correlations were also examined at the subscale and scale level as well. The intraclass correlation is interpreted much like a reliability coefficient with values closer to 1.0 indicating good agreement and low values indicating poor agreement (Winter & Calsyn, 2000).

Agreement between data sources. As discussed in Chapter 4, the information for fidelity ratings often can be obtained from several sources, such as program staff, direct observation, and agency records. When two or more sources of fidelity ratings are used, then agreement between sources can be determined, using the same statistical methods as used for assessing agreement between raters (i.e., intraclass correlation and kappa). Another (low-tech) method for assessing agreement between two sources is to calculate the Pearson correlation and t test across sites, as illustrated in Table 5.4. The Pearson correlation indicates the concordance in relative ranking of sites, while the t test determines if there is any systematic bias in the ratings. Assessment of agreement can be done at the item, subscale, and total scale level. We recommend examining all these levels, depending on the goal of the evaluation. The item level analysis helps to pinpoint which items are

most troublesome, whereas assessment of the total scale is most useful to determining globally how well the sources agree.

Table 5.4. Comparison of Ratings on Team Leader and Case Manager Responses to Ratings of Chart Data for Selected DACTS Items from Zahrt et al., 1999).

	Team Leader	Chart	Team Leader vs. Chart	
Variable	**M (SD)**	**M (SD)**	**t test**	**r**
Small Caseload	4.6 (0.5)	4.9 (0.3)	-1.96 t	.41
Team Approach	3.8 (0.8)	3.4 (1.2)	1.31	.58 t
Team Meeting	4.1 (0.8)	4.4 (0.5)	-2.00 t	.78*
Continuity of Staffing	1.9 (1.2)	2.1 (1.2)	-0.48	.29
Staff Capacity	3.8 (0.8)	4.2 (0.8)	-2.53*	.80**
Intake Rate	5.0 (0.0)	4.9 (0.3)	1.00	--
In-Vivo Services	4.2 (0.6)	4.1 (1.1)	0.25	-.03
No Dropout Policy	4.5 (0.7)	4.1 (0.7)	1.18	-.11
Intensity of Services	4.6 (0.7)	3.6 (0.7)	2.37*	-.82**
Frequency of Contact	3.3 (0.5)	2.6 (0.8)	3.28**	.60 t

NOTE: Scores range from 1 to 5, with 1 = not implemented and 5 = fully implemented.

t p <.10 * p <.05 ** p <.01 *** p <.001

One study examining source agreement was conducted by Vogler (1998), in her study of supported employment programs. Vogler interviewed multiple sources in 20 programs, including program supervisors, either one or two employment specialists (ES1 and ES2), and follow-along specialists, as shown in Table 5.5. Her fidelity instrument consisted of 3 subscales: Human Resources, Organization, and Services. Agreement between the supervisors and the employment specialists was very good for the Human Resources and Organization subscales, but poor for the Services subscale. Agreement with the follow-along specialists was poor across subscales, as reflected by the reliability coefficients. Vogler speculated that follow-along specialists in a supported employment program were poor informants, because they were not involved in the early stages of the rehabilitation process. Two general conclusions are that (1) sources generally agree better on low inference questions, and (2) sources who are more involved in the day-to-day activity are better informants.

Table 5.5. Intraclass correlations for ratings made with different data sources (from Vogler, 1998)

Scale	Supervisor vs ES1 (N=20)	Supervisor vs ES2 (N=9)	ES1 vs ES2 (N=9)	Super. vs Follow-Along Specialist (N=6)	Follow-Along Specialist vs ES (N=6)
Human Resources	.85	.91	.91	.44	.17
Organization	.77	.74	.89	.57	.42
Services	.38	.47	.71	.45	.04
Total	.65	.63	.90	.16	.44

NOTE: Super = Supervisor; ES1-2 = Employment Specialist 1 & 2

Zahrt et al. (1999) also examined agreement between ratings based on interviews with different sources. This study used the DACTS to rate 10 newly-developed ACT programs. Data sources included interviews with team leaders and case managers and 10 randomly selected charts. As shown in Table 5.4, Pearson correlations between team leaders and chart data were mixed. The largest discrepancies were found on items relating to service intensity, with chart data suggesting less intensive services than reported by either the team leader or the case manager.

The obvious next question if there is poor agreement among sources is how to proceed in measurement. The options are to depend on a single source, believed to be the best, combine information from different sources, using a fixed decision rule (e.g., take the average rating), use clinical judgment on a case-by-case basis, or discard items deemed unreliable. We do not have a perfect solution to this question. However, Winter and Calsyn (2000) offer a thoughtful discussion of these issues.

Recommendations for assessment of reliability

As a bare minimum, developers of a new fidelity scale should conduct an item analysis and examine internal consistency and inter-rater agreement. An example of an item analysis is given in Table 5.2, which was part of the psychometric study of the IPS Fidelity Scale (Bond et al., 1997a). See Appendix for more details on the scale. This item analysis shows the distribution of item responses and the inter-rater agreement. Not shown in this table are item-total correlations and subscale internal consistency coefficients for Staffing and Organization subscales.

Validity

Validity refers to the degree to which the scale measures what the researcher purported it to measure. If a scale measured something different than what it is intended to, any inferences made about the responses to that scale will be erroneous. An example of an invalid fidelity scale would be a measure that always rated larger programs as higher in fidelity, regardless of the actual interventions provided. It is important, therefore, that we provide evidence regarding the validity of our measures.

Traditionally, five types of evidence have been used to support the validity of measures. Face validity is the least empirically based, and refers to whether the items appear to be tapping into the construct that they should. Face validity is typically more critical for the respondent than the researcher. Content validity refers to the content of the items and how representative they are of the larger pool of content. Construct validity refers to whether the scale measures the particular construct that it was intended to measure. This is typically indicated through evidence of convergent and divergent validity. Criterion-oriented validity refers to methods to validate a measure through external criteria and includes predictive validity. Finally, predictive validity reports the relationship between the measure and an anticipated criterion.

Face validity

Face validity evidence is typically more critical in terms of respondent perceptions or others' beliefs about the test, than empirical evidence relevant to the researcher. In short, face validity refers to the perception that the items seem to measure what they should. It is important that the respondents feel comfortable answering questions and to the degree that items seem appropriate, it is likely respondents will not react negatively toward the experience. Although the credibility of the scale from the respondent's perspective is important, face validity is also important for establishing credibility among psychiatric rehabilitation practitioners, consumers, researchers and other stakeholder groups.

Content validity

Content validity evidence refers to the appropriateness of the content or material in the test as a representation of the relevant content. A content valid test is one that can show that the items chosen are representative of that pool. The evidence for this kind of validity actually comes from documentation and planning that occur prior to the development of the test. In Chapter 3 we gave examples of studies that sought to establish content validity through consensus panels (Marty et al., under review; McEvoy et al., 1999; McGrew & Bond, 1995; Schaedle & Epstein, 2000). Although these studies occurred in an earlier stage of instrument development, there is no reason that evaluators could not present the actual fidelity measures to an expert panel for review, thereby making the content validity argument even stronger.

Construct validity

The basic definition of validity includes the assumption that we are measuring the construct we think we are. A common way to assess this is through a multi-trait multi-method approach (Campbell & Fiske, 1959). In this approach we use several different methods (i.e., self-administrated scale, interviews with various constituents) to assess a variety of constructs. Given that we have suggested using multiple data sources for assessing fidelity, this is a likely way to assess the validity of the fidelity scale. As noted above, it is likely that there will be some level of disagreement for some items. We cannot guide the researchers in identifying a "best" source, however, we would encourage the reconciliation of differences between sources through follow-up interviews.

A second way that construct validity evidence is documented is by correlating scores from an existing measure and a new measure. For instance, if a fidelity measure is constructed for a supported employment program, the responses to this scale should correlate with other measures of fidelity for supported employment which follow the same specific model (convergent validity). Similarly, the responses for a fidelity measure

constructed for a supported employment program should not be highly related to scores for a fidelity measure constructed for an ACT program (divergent validity).

Known groups validity

One common purpose of fidelity measures is to demonstrate that programs that follow a particular model are distinctive. For example, Bond et al. (1997) piloted the IPS Fidelity Scale in 27 sites, including 9 IPS programs, 11 other supported employment programs, and 7 other vocational rehabilitation (VR) programs. As shown in Table 5.6, the scale discriminated between IPS and the other VR programs. In further analyses (not shown), IPS and the other SE programs differed on items relating to integration with mental health services and zero exclusion admission criteria.

Table 5.6 Agreement Between Program Label and Fidelity Rating (from Bond et al., 1997)

	Fidelity Rating			
Program Label	Consistent with IPS 66 - 75	Partially Consistent with IPS 56 - 65	Not IPS < 55	Total
IPS	8	1	0	9
Other SE	2	8	1	11
Other VR	0	2	5	7
Total	10	11	6	27

In another example of known-groups validity, Teague et al. (1998) piloted the DACTS in 50 programs representing four distinct types of service models: ACT, intensive case management provided by the Veterans Administration, outreach programs for people who were homeless and mentally ill, and traditional case management. The DACTS discriminated across the four types of case management, consistent with predicted order of similarity to ACT. However, a subsequent study comparing DACTS ratings in a sample of 18 outreach programs for homeless people with mental illness found nonsignificant differences between programs subscribing to nominally different program models (Johnsen et al., 1999). Thus, more work is needed to determine the capacity of the DACTS to make fine-grained discriminations.

Predictive validity

We typically anticipate that if a measure is valid, than it should be predictive of some criterion. For instance, with a fidelity measure, we anticipate that a program which has been well implemented and scores highly on the fidelity scale should also have positive outcome criteria like decrease in hospital stay for case management programs, longer tenure in jobs for vocational programs, and increased satisfaction with recreational activities for drop in centers. These relationships are typically tested with a correlation coefficient to provide predictive validity evidence. In this case, the assumption is that if our scale truly measures program fidelity, it should be highly and positively related to the outcome measures.

In Table 5.7 we present an example of predictive validity at total scale, subscale, and item level in the study conducted by McGrew et al. (1994). Overall, the scale was highly predictive, with the Staffing and Organiza-

tion subscales significantly correlated with outcome, whereas the Services subscale was not. Several items emerged as significant predictors, as shown by the rank ordering in Table 5.7.

Table 5.7. Correlations Between Ratings on ACT Fidelity Items and Reduction in Days Hospitalized for 18 ACT Teams (adapted from McGrew et al., 1994)

Shared caseloads	.65**
# Total contacts	.59**
24-Hour availability	.55*
Nurse on team	.49*
Daily team meetings	.49*
Coordinator provides services	.46
Team size	.35
# In vivo contacts	.31
Psychiatrist on team	.28
Time-unlimited services	.28
% of contacts in community	.21
Client:staff ratio	.19
Separate site	.18
Total hours of contacts	.16
Team primary therapist	.06
Hours in vivo contacts	-.03
% Hours of office contacts	-.11
Staffing Subscale	.54*
Organization Subscale	.56**
Service Subscale	.33
Total Fidelity Scale	.60
*p<.05, ** p<.01	

Although it may seem a simple decision, the choice of the criterion is often a difficult one. Many researchers spend a great deal of time developing and modifying the fidelity measure, but do not give as much attention to the development of the criterion measure. This issue has been termed the "criterion problem" (Austin, 1992). Often, we find that the criteria used are chosen on the basis of convenience (e.g., existing reports.), rather than well-developed measures. If the criterion measure is poor or unreliable, the correlation between the predictor (i.e., fidelity measure) and the criterion (i.e., performance measure) will be low. Thus, a low predictive correlation may be obtained which could lead to the conclusion that (1) the measure is not valid, or (2) the criterion measure is unreliable, or (3) both measures are unreliable.

Another issue to consider is the level of measurement of the criterion. In assessing the predictive validity of a fidelity measure, the focus of interest is at the program level, rather than the individual level. Thus, when examining the predictive validity of a fidelity measure, the criterion of interest should be at the program level. Many times, this will require that individual level data (satisfaction measures, hospitalization rates) are aggregated in some fashion. Methods of aggregation most often used for fidelity measures include mean rates (e.g., employment rates) and mean effect size. Another method to capture a group level construct with

individual level data is through the use of a statistic termed rwg. In this calculation, the statistic represents the within-group consensus of the individual responses as the operationalization of the higher-level group construct. The types of constructs that may be appropriate for this analysis include perceptions of climate measured through Moos' environment scale (Moos, 1974a) and the Program Environment Scale (Burt et al., 1998). For a full review of this procedure, please see James, Demaree, & Wolf (1995).

The choice of criterion measure should be driven by the empirical literature. For instance, the ACT program has a strong impact on hospitalization and a moderate effect on independent living, but not a strong effect on other outcomes (Latimer, 1999b; Mueser et al., 1998). Similarly, vocational programs affect employment outcomes much more than other criteria (Bond et al., 1997b).

It should be noted that it is often difficult to develop conceptually clear measures of the desired outcomes. For instance, in a vocational program one of the criteria might be adequate income. Obviously, the most direct actual measure for this criterion is hourly wages. Even this, however, is an imperfect measure, because it will be influenced by a number of other factors: area in the country, unemployment rate in area, etc. These factors are examples of criterion contamination, which suggests that the measure is contaminated with another construct (e.g., unemployment rate, client functioning level).

Other aspects of the criterion measure to consider are range restriction, time of measurement, and some other third variable problem. First, if a criterion measure has restriction of range, this will lower the correlation between it and the predictor measure (i.e., fidelity scale). Additionally, the longer the time period between the predictor and criterion, the lower the correlation is likely to be. For instance, if the criterion is not measured for several months following the fidelity measurement, then program may change in the interim. Finally, if the respondent consulted in completing the fidelity ratings also is responsible for the ratings on the criterion measure, the correlation between the two measures may be inflated due to method variance (Campbell & Fiske, 1959).

Step 14 Determine Scoring and Weighting of Items

We began Chapter 3 noting our recommendation for simple linear additive scales. Such scales typically use unit weightings, although other weighting schemes could also be used. In principle, weighting items based on their importance or criticality to the program might produce more valid data. For example, a case management program that does not make home visits cannot be considered an ACT program, regardless of how well it satisfies other criteria. Following this reasoning, an item regarding in vivo visits should be weighted more heavily than any other item. A systematic way to capture perceived importance could be to use expert surveys (e.g., McGrew & Bond, 1995). Factor scores derived from factor analyses is another. Researchers have experimented with such schemes in their fidelity scale scoring, although significant improvements in the predictive validity have not consistently been found between weighted and unweighted solutions (Johnsen et al., 1999; McGrew et al., 1994). Cascio (1998) suggested that in most cases, unit weighting (i.e., each item receives a weight of 1) may not only be appropriate, but may be preferred (For a full review, see Dawes & Corrigan, 1974).

Chapter 6. Conclusions and recommendations

Fidelity is clearly a useful tool in both research and practical settings. In the last few years, progress has been made in the theoretical conceptualization of fidelity as well as the acceptance of fidelity measurement in the scientific community as a necessary tool in improving research. Although discussion of the practical uses of fidelity is lacking in the literature, we are seeing increasing demands for the measurement of adherence to program standards, not only from the scientific community, but also from a variety of stakeholders involved in funding, providing, and receiving psychiatric rehabilitation services. The use of fidelity measures in practical settings can have salutary effects on quality improvement. A litmus test that a fidelity measure has become institutionalized occurs when individual fidelity checklist items become the basis for policy discussions around what the standards should be. Insofar as these discussions become centered on specific data and empirical criteria, these debates represent a constructive advance over policy driven by politics.

Researchers have begun to develop standards for the development and use of fidelity measures in the field of psychiatric rehabilitation (Calsyn, 2000). Some also have begun to use fidelity measures as a way to improve experimental studies in psychiatric rehabilitation. As seen in the Appendix, a number of efforts are under way to develop fidelity measures spanning several areas of psychiatric rehabilitation. We have attempted to capture that process here, providing a guide with which to begin to develop new measures, or to improve existing ones.

Although the development of fidelity measures has been welcomed in most research and clinical arenas, there are still many questions left unanswered. As discussed in Chapters 3, 4, and 5, the process of developing a scale offers a myriad of choice points. Throughout this toolkit, we have offered recommendations for measure development. Although we have sought to bring empirical data to bear on method questions, we harbor no illusions that this toolkit offers the final answer to critical questions. Many challenges remain in the process of developing measures in the field of psychiatric rehabilitation. We highlight several particularly important issues for consideration below.

One key challenge to fidelity assessment is lack of clarity in the model itself. As we have discussed, psychiatric rehabilitation is composed of eclectic approaches, many of which are difficult to operationalize. Other approaches lack consensus on key components. Without uniform guidelines that are clearly outlined, for example, in practice manuals, fidelity to a model cannot be assessed. In addition, some programs are evolving, and it may be too early in their development to create a fidelity measure. As practice manuals become more widely used, the need and ability to create fidelity measures will greatly increase.

Even when models are clearly defined and have agreed upon dimensions, some elements may be difficult to quantify. For example, when assessing aspects of service provision, frequency of contact and length of contacts are relatively straightforward. However, it is more difficult to assess what happens during a contact, how service is provided, and the overall quality of the contact. In addition, there may be very unique services that a program provides which are not assessed by a fidelity measure. These unique elements often go unaccounted for when programs are assessed by structured fidelity interviews. For example, Waltz et al. (1993) discuss techniques to assess competence in the psychotherapy domain that may prove useful for fidelity measurement in psychiatric rehabilitation as well.

When developing any new measure, it is important to demonstrate reliability and validity. As discussed in Chapter 5, many traditional indicators of reliability and validity are less applicable to fidelity measurement. One of the most widely used indices of reliability is internal consistency. However, fidelity measures are multidimensional, even within subscales. Because of the breadth of dimensions covered in fidelity assessment, too few items measure the same construct to do a meaningful examination of internal consistency. As discussed by Winter and Calsyn (2000), measures of internal consistency may not be appropriate for fidelity indices, at least for some program models. Other forms of reliability, for example stability and inter-rater agreement, may be more suitable.

Fidelity measurement is also challenged by sampling issues. Because fidelity is assessed at the program level, we are limited in the sample sizes we have available. Small sample sizes make some analyses problematic (e.g., unstable correlations), limit the power to detect differences in programs or relationships with outcomes, and hamper generalizability of findings. A related sampling issue is restriction in range. If the sample is comprised of programs that are supposed to be following a particular model and the programs are well implemented, there will be little variability on the items and fidelity scores will tend to cluster at the high end of the scale. This restriction in range affects correlational statistics, so that items or subscales may not correlate strongly with other variables, even if in reality the relationships are very strong. In order to overcome this, it is advisable to assess wide range of programs (Teague et al., 1998).

We recommend the inclusion of multiple perspectives in fidelity assessment (i.e., multiple sources of data, multiple methods for data collection, and multiple raters). However, this poses many challenges for fidelity assessment. In terms of multiple sources, we believe some sources may provide more reliable information for certain dimensions and not for others. That is, some informants may be more knowledgeable about specific aspects of their program and some methods (e.g., interview vs. chart review) may be more accurate for some dimensions. Thus, for some items, agreement between sources or methods may be low. Similar disagreements between sources of information have been noted in the psychotherapy literature, for example, process notes and videotaped therapy sessions may result in different ratings of fidelity (Waltz et al., 1993). Because we expect some disagreement between sources and between methods, establishing agreement between raters is of utmost importance.

The use of multiple perspectives also creates problems of how to combine the information. Should responses from different sources or raters be weighted equally? Should some information be discarded if it appears to be inaccurate? If so, how should such decisions be made? This particular challenge reflects a longstanding debate in psychology between the accuracy of clinical judgment and actuarial methods (Meehl, 1954). Ideally, empirically based (i.e., actuarial) methods could be developed; however, we would still need to rely on judgment about which information is collected.

Another challenge to fidelity measurement concerns the timing of the assessment. In new programs, we expect some start-up time before a program is fully operational and running well. Using fidelity measures during these stages can help guide development of programs by providing concrete feedback in areas that need attention. However, if we try to assess the stability of our fidelity measure during this phase, it can result in low stability, especially if there is a long delay between assessments. In this case, it is likely that actual changes are affecting stability, rather than error in the measure. Similarly, if the purpose of fidelity measurement is to compare a program to existing programs following the same model, the developmental phase of the

program will affect the conclusions of the fidelity assessment. For example, a new program being compared to a well-established program may be an unfair comparison if the timing of the fidelity assessment is too early or if the context of the developmental phase of the program is not considered. Thus, close attention to the timing of fidelity assessment is important.

As seen in the challenges listed above, developing valid and reliable fidelity measures is a complicated process. However, we believe the potential usefulness of fidelity measures outweighs the difficulties faced in the developmental process. The dilemmas above need further study and should not be ignored in discussing and developing fidelity measures.

In summary, for those developing fidelity measures, it is important to clarify the use of the fidelity scale before beginning development. The type of use will dictate much of the process. We also recommend a careful study of the model to be measured, being certain to clarify the model dimensions. In developing a measure, it is important to be flexible and willing to try different methods. We have made recommendations for the ideal ways to develop measures; however, these will not apply to all situations. A thorough pilot of the scale will ensure that many of the more crude problems are remedied before applying the scale in the field.

As seen in the Appendix, most fidelity measures in the psychiatric rehabilitation field are rudimentary. We have our work cut out for us if we want to pursue an agenda using fidelity measures to improve research, and to improve psychiatric rehabilitation services. More research is needed to examine the hypothesis that increased fidelity to a program model results in improved client outcomes. Certainly, we would expect this hypothesis to be supported most often for program models that are validated through randomized controlled trials. In addition, the use of fidelity measures as evaluation tools in practical settings requires further study to identify the processes that benefit organizations and programs in improving services. We believe that we should apply the lessons from the psychotherapy fidelity literature, as well as the broader literature on measurement. Although fidelity measurement is no panacea, it can help in both the research and practice of psychiatric rehabilitation.

Appendix

Table of Instruments in Use

	Psychiatric Rehab Practicesa	PESa,e	PRESa,e
References	Anthony, Cohen, & Farkas, 1982; Farkas et al., 1988; & Fishbein, 1988	Burt et al., 1998	Evans et al., 1998
Item generation	Anthony et al., 1982	literature on CSP, existing scales, advisory group of program directors	through literature, Connect98 initiatives, and expert consultation
Instrument length	54 items	97 T/F items	34 items
Reliability/ validity sample	50 partial care programs in NJ	221 clients in 22 programs	12 (pilot) and 74 (full-scale evaluation) mental health centers in IL
Data sources	interviews with clients, chart review, review of manuals and 1-day site visits	client interview	interview with programs director or staff, observation of program
Time to administer	15-30 minute interview with client	25-minute checklist interview with client	45 minute interview
Subscale structure	8 subscales	3 domains/24 subscales	3 subscales
Internal consistency (Cronbach's alpha)	not assessed	range =.61- .93	total scale =.71; subscales =.78.67.76
Stability/sensitivity to changes	sensitive to changes over time		some evidence for stability
Agreement		not applicable	average agreement between raters = 78%
Validity	not reported	evidence for concurrent validity & known-groups validation	evidence for concurrent validity

Other information		very comprehensive scale	3 items with ceiling effects 2 with floor effects

	Clubhouse Normse,a	Principles of Psychosocial Rehabilitation Scalea	DACTSb
References	Bond, Vogler, & Irmscher, 1994; Macias & Jackson, 1993	Lucca, 1998	Teague et al., 1998; Winter & Calsyn, 2000; Zahrt et al., 1999; Johnsen et al., 1999; & Salyers et al., 1998
Item generation	researcher generated	adapted from Bachrach (1992) & Cnaan et al. (1998)	existing scales, expert opinion
Instrument length	18 item checklist	20 items	26 behaviorally-anchored items
Reliability/ validity sample	158 clients	24 psychiatric rehab programs in New England	50 case management programs
Data sources	client self-administered	interview with program staff	team leader, case managers, chart reviews
Time to administer	10 minutes		one-hour interview with team leader
Subscale structure	6 subscales	none	5 subscales
Internal consistency (Cronbach's alpha)	subscales =.84.69.69.43.71.69	total scale=.88	total scale =.92; subscales =.87.87.77.74.77
Stability/sensitivity to changes	some evidence for stability	not reported	low stability
Agreement			mixed agreement between data sources
Validity	some evidence for known-groups validation	good face validity	evidence for predictive validity & known-groups validation, good face validity
Other information			no variation on 3 items covers only portion of elements experts deem critical used as monitoring tool in statewide ACT evaluation

	IPS Fidelityc	QSEISc	SISTEMc
References	Bond et al., 1997, 1999	Bond, Picone et al., 1999	Vogler, 1998
Item generation	from model developers and treatment manual	IPS scale, advisory board, best practices	developed for use in Indiana, expansion of IPS scale
Instrument length	15 behaviorally-anchored items	33 behaviorally-anchored items	35 behaviorally-anchored items
Reliability/validity sample	27 vocational programs (9 IPS, 11 other SE, 7 other VR)	32 SE programs in NJ and Kansas	24 sites
Data sources	interviews with staff, site visits, objective records	team leader interview	staff interviews
Time to administer	one-hour interview	1.5 hour interview	1 hour interview
Subscale structure	3 subscales	4 subscales	4 subscales
Internal consistency (Cronbach's alpha)	total scale =.92; subscales =.72.65.90	total =.51, subscales =.74.60.74.62	total =.59, subscales =.48.56.05.60
Stability/sensitivity to changes		not assessed	
Agreement	interrater reliability (average >.80)	84% agreement between raters, poor agreement between state administrators & interviewer	ICC between raters =.97, agreement between sources variable
Validity	evidence for concurrent and predictive validity and known-groups validation	evidence for content, predictive and concurrent validity, and known-groups validation	evidence for known-groups validation, content & concurrent validity
Other information	evidence for ceiling effect (Scale values 4 & 5 were used more than 70% for all items) used in research/ monitoring projects	possible ceiling effects data collection continues	intended for use in Indiana ratings positively skewed

	DPA[c]	Ideologies of Care[d]	Fidelity Assessment for the Center for Mental Health Service Housing Initiative [d]	Member Involvement Scale [e]
References	Rollins et al., 2000	Heaney & Burke, 1995	Lassiter (personal communication, 1999)	Mowbray, Robinson, & Holter, 1999
Item generation	scale under development, see the below for more information	pilot to generate items	scale under development, see the below for more information	meeting with JMHO staff and consumers in other drop-in centers
Instrument length		14-item checklist, Likert scale		10 items
Reliability/ validity sample		staff in 269 group homes in Michigan		33 drop-in centers in Michigan
Data sources		residential staff		phone interview with director of the center
Time to administer				1 hour
Subscale structure		2 subscales		none
Internal consistency (Cronbach's alpha)		subscales =.79.80		Total=.82
Stability/sensitivity to changes				
Agreement				
Validity		not reported		evidence for known-groups validation
Other information				

	Clubhouse Fidelity Index[e]	Therapist Fidelity Evaluation for Skills Training Checklist	Therapist Fidelity & Competency Scale[g]
References	Lucca, 2000	Wallace et al., 1992	Mueser, Gingerich, & Rosenthal, 1994
Item generation	theory and empirical literature	manual	researchers generated items (fidelity subscales) and modified an existing scale (competency subscales)
Instrument length	15 yes/no items	170 items on checklist	9 dichotomous items for each of fidelity subscales and 4 items for competency subscales
Reliability/ validity sample	22 PSR programs in Connecticut	7 treatment facilities	8 families
Data sources	interview with staff	skills training group observed	family sessions were audio taped and rated
Time to administer	15-30 minutes	length of skills training group	
Subscale structure	none	not assessed	2 fidelity and 4 competency subscales
Internal consistency (Cronbach's alpha)	total scale =.75	3 modules: .87.64.73	Fidelity=.93.87 Competency=.82.74.69.86
Stability/sensitivity to changes		not reported	
Agreement		agreement between raters =.88 (kappa)	ICC=.93 (BFT-Fidelity) and.87 (EFT-Fidelity). ICC= Structure (.82), Relationship (.74), Difficulties (.69), and Global (.86)
Validity	some evidence for concurrent validity and known-groups validation		Evidence for known-groups validation
Other information	focus on vocational dimension only	closely follows manual	

[a] General Psychiatric Rehabilitation Scales
[b] Case Management Scales
[c] Vocational Program Scales
[d] Residential Program Scales
[e] Drop-in Center Scales
[f] Skills-training Scale
[g] Family Psychoeducation Scale

Instruments in Use

* Scale described F = Fidelity Scale I = Implementation Scale

I. General Psychiatric Rehabilitation/Program Environment Scales

- *Psychiatric Rehabilitation Practices (Farkas et al., 1988; Fishbein, 1988) F
- *Program Environment Scale (PES) (Burt et al., 1998; Burt & Hargreaves, 1997) I
- *Psychiatric Rehabilitation Environment Scale (PRES) (Evans et al., 1998) F
- *Clubhouse Norms (Macias & Jackson, 1993) I
- Community Program Philosophy Scale (Hargreaves et al., 1998; Jerrell & Hargreaves, 1991) I
- Community Oriented Program Environment Scale (Moos, 1974a) I
- *Principles of Psychosocial Rehabilitation Scale (Lucca, 1998) F

II. Case Management Scales

- *Dartmouth Assertive Community Treatment Scale (DACTS) (Teague et al., 1998) F
- Latimer ACT Fidelity Scale (Latimer, 1999b) F
- Strength Case Management Scale (Rapp, 1999) F

III. Vocational Program Scales

- *Individual Placement and Support (IPS) Fidelity Scale (Bond et al., 1997a) F
- *Quality of Supported Employment Implementation Scale (QSEIS) (Bond et al., 1998c) I
- *Scale for the Indiana Supported Employment Model (SISTEM) (Vogler, 1998) F
- Standards of Excellence for Employment Support Services (Wood & Steere, 1992) F
- Procedural Components of Supported Employment Programs (McDonnell, Nofs, Hardman, & Chambless, 1989) I
- *Diversified Placement Approach (DPA) Fidelity Scale (scale under development, Rollins, Bond, Salyers, Resnick, Dincin, McCoy, Kinley, Shimon, Marcelle, Fraser, & Forman, 2000) F

IV. Residential Program Scales

- *Ideologies of Care (Heaney & Burke, 1995) I
- *Fidelity Assessment for the Center for Mental Health Services Housing Initiative (scale under development, Lassiter, 1999) F

V. Drop-In Center Scales

- See Psychiatric Rehabilitation Environment Scale under Program Environment Scales
- See Clubhouse Norms under Program Environment Scales

- *Member Involvement Scale (Mowbray, Robinson, & Holter, 1999) F

VI. Clubhouse Scales

- *Clubhouse Fidelity Index (Lucca, 1998) F
- See Clubhouse Norms under Program Environment Scales
- See Program Environment Scale under Program Environment Scales

VII. Skills Training Scales

- Skills Training
- * Therapist Fidelity Evaluation for Skills Training Checklist (Wallace et al., 1992) F

VIII. Family Psychoeducation Scales

- *Therapist Fidelity and Competency Scale (Mueser et al., 1994) F

IX. Supported Education Scales

- None located as of 1/4/00

I. General Psychiatric Rehabilitation/Program Environment Scales

Psychiatric Rehabilitation Practices

References. (Anthony et al., 1982; Farkas et al., 1988; Fishbein, 1988)

Item generation. Principles and case examples found in Anthony et al. (Anthony et al., 1982)

I**nstrument length**. 54 items.

Reliability/validity sample. 50 partial care programs in NJ (Fishbein, 1988)

Data sources. Documentation of Policy and Procedures Manuals; Record review; Interviews with 55 clients. 1-day site visits by two state monitoring agencies. Client interviews were 15-30 minutes in duration.

Subscale structure. 8 subscales: Treatment goal described, Skill strength assessed, Skill deficit assessed, Resource strength assessed, Resource deficits assessed, Master plan described, Interventions identified, Client involvement noted.

Reliability. Stability of the scale- Mean percentage compliance was 38% at pre-test and 80% at post-test.

Validity. Not reported

Recommendations. This methodology appears to be most suited for pointing out discrepancies between the label of psychiatric rehabilitation and actual practice (Farkas et al., 1988). It is useful for documenting the paperwork compliance within an agency. The change over time suggests that it is a useful monitoring tool. The key question is the relationship between compliance and client outcomes. Because the recommended methodology requires daylong site visits, it is expensive.

Program Environment Scale (PES)

References. (Burt et al., 1998)

Item generation. Began with 10 broad categories of program life found in the literature on community support programs: Growth and Enhancement; Belongingness; Peer Support; Helpfulness of services; Community Resources; Staff cohesion; Staff attitudes toward consumers; Empowerment in choice of treatment; Governance; Staff accessibility. Researchers examined existing scales, scales-in-progress, theoretical literature, consulted with an advisory group of program directors and experts.

Instrument length. 97 true-false attitudinal items.

Reliability/validity sample. Pretest with 121 clients in 12 randomly selected programs near Washington, DC; final field test with 221 clients in 22 randomly selected programs across U.S. Final sample included 5 certified clubhouses, 8 day treatment /partial hospitalization programs, 6 psychosocial rehabilitation programs, 2 social clubs, and 1 other.

Data sources. 25-minute checklist is given in an interview with clients; staff version also developed.

Subscale structure. Three domains: Atmosphere/interactions; Client empowerment/staff-client equality; Service components. 24 subscales, of which 23 met 5 of 8 psychometric criteria for internal consistency and discriminant validity. The item convergence criteria were (p. 865): (a) Cronbach's alpha greater than or equal to.7; (b) ratio of variance of item with the least variance to item with the most variance less than or equal to 2.0; (c) minimum item-scale correlation (with item removed) =.3; (d) difference between highest and lowest item-to-total correlation less than or equal to.20; (e) subscale items produce a single factor, indicating unidimensionality; (f) loadings on the first unrotated factor great than or equal to.6; (g) difference between the highest and lowest factor loadings less than or equal to.20. The item discriminant validity criteria were: (a) scale significantly discriminate among programs; (b) subscale items correlate with own subscale at least two standard errors higher than they do with any other subscale.

Reliability.

Internal consistency (Cronbach's alpha). Range from .61 to .93 for individual subscales

Validity.

Known-groups validation. A sample of clubhouses were associated with higher ratings on the following subscales: staff-client respect, availability of good touch, all three empowerment subscales, importance of work, and negatively associated with medications and substance abuse treatment. Day treatment/partial hospitalization programs were positively associated with subscales of medications and substance abuse treatment, and family activities.

Concurrent validity. Associations between PES and information from Program directors were strong

Other information. Comprehensiveness is strength of this instrument.

Recommendations. Authors consider this an implementation scale and not a fidelity scale, with "neutral" item. The authors had originally intended the PES to cover all types of community -based programs. However, many items proved to be not suitable for ACT. Another concern with this scale is that only 116 of 221 respondents completed all the items.

Psychiatric Rehabilitation Environment Scale (PRES)

References. (Evans et al., 1998)

Item generation. This scale is intended to measure how an agency environment fosters consumer empowerment, advocacy, independence, and involvement. Three sources were used for the development of the content and structure of these instruments. First, the program component descriptions in guidelines developed by the Illinois Office of Mental health were used as the basis for each instrument (OMH, 1997). Second, experts in the field of PSR evaluation were contacted. Finally, we conducted an extensive review of the literature, with particular attention to the literature on fidelity and scale development in the area of PSR.

Instrument length. 34 items.

Reliability/validity sample. 12 mental health centers in Illinois were piloted in 1998. 74 mental health centers were used for full-scale evaluation in 1999.

Data sources. Interview with program director or other program staff familiar with peer support programming. Potentially, the first 24 items could be administered over the phone, however, the scale has not been tested in this manner. The last 10 items require an on-site visit. Interview takes 45 minutes.

Subscale structure. A factor analysis yields 3 factors of Activities and Structure of Program (10 items), Consumer Empowerment (14 items), and Accessibility (3 items). Consumer Empowerment and Activities and Structure were correlated with each other (r=.37). The Accessibility factor was not significantly correlated with the other two factors.

Reliability.

- Internal consistency (Cronbach's alpha). Total score (.71). Activities and Structure of Program (.78), Consumer Empowerment (.67), and Accessibility (.76).
- Agreement between interviewers. Percent agreement between raters ranged from 50% to 100%, with overall agreement averaging 77.7%.
- Stability of the scale. For 12 mental health centers rated in pilot (1998), the correlation between two sets of rating in 1998 and 1999 was .78.

Validity.

- Concurrent validity. Global interviewer ratings were correlated significantly with the PRES (r =.91 and r =.92, respectively for the two interviewers).

Other information. Evidence for floor/ceiling effects. All but 5 items showed variability across the entire 5-point scale. There were 3 items with ceiling effects (no agencies received score of 1 or 2) and 2 with floor effects (no agency received score of 4 or 5).

Recommendations. The PRES proved to be a face valid tool for describing general adherence to psychosocial rehabilitation principles, related to the peer support component of CONNECT98. We recommend this scale for use by state mental health agency staff and/or CMHC staff for monitoring purposes.

Clubhouse Norms

References. (Bond, Vogler, & Irmscher, 1994; Macias & Jackson, 1993)

Item generation. Investigator-generated checklist intended to measure member-staff relationships in clubhouse programs (Macias & Jackson, 1993).

Instrument length. 18 items

Reliability/validity sample. 158 clients attending one of 3 psychosocial rehabilitation programs (PSR sample) and 113 clients attending a community mental health center (CMHC sample) (Bond et al., 1994).

Data sources. Completed by clients attending either a psychosocial rehabilitation center or a mental health center. It takes 10 minutes to complete questionnaire.

Subscale structure. Six 3-item scales, defined a priori

Reliability.

- Internal consistency (Cronbach's alpha). Member-Member Affiliation (.84), Member Influence (.69), Staff Cohesion (.69), Member-Staff Affiliation (.43), Member-Member Help (.71), and Member Mastery (.69).
- Stability of the scale. In the Bond et al. (1994) study, the scale was re-administered 4 months after the initial administration. The data were not analyzed statistically, but graphically the mean levels appear relatively stable.

Validity.

- Known-groups validation. The PSR sample was significantly higher than the CMHC sample on two of the scales (Member-Member Affiliation and Member-Staff Affiliation). In each case, the higher ratings were in the predicted direction, with a greater sense of camaraderie between clients and between clients and staff. There were no differences on the other four subscales.

Recommendations. May overlap with the PES, but the scales seem to have decent face validity. The main drawback to this instrument is the lack of psychometric information.

Principles of Psychosocial Rehabilitation Scale

References. (Lucca, 1998)

Item generation. Adapted from Bachrach (1992) and Cnaan et al. (1988), Lucca lists 10 core values: individualized rehabilitation, environmental focus, restoration of hope, focus on strengths, vocational potential, continuity of care, normalization, comprehensiveness of care, consumer involvement, and staff de-professionalization.

Instrument length and administration. 20 items

Reliability/validity sample. 24 psychiatric rehabilitation programs in a New England state

Data sources. Interviews with program staff

Subscale structure. None

Reliability. Internal consistency (Cronbach's alpha) Total score (.88)

Validity. Not reported.

Recommendations: Scale has good face validity, may be suitable as a brief scale for measuring program philosophy.

II. Case Management Scales

Dartmouth ACT Scale (DACTS)

References. (Teague et al., 1998; Johnsen et al., 1999; Salyers et al., 1998; Winter & Calsyn, 1999; (Zahrt et al., 1999)

Item generation. Adapted from prior ACT fidelity scales (McGrew et al., 1994; Teague et al., 1995). "The DACTS addresses some limitations in scaling, explicitness, and comprehensiveness of the earlier instruments. The majority of criteria and anchors in the new instrument were adapted from variables used as indicators for the 13 dimensions reported in Teague, et al. (1995), and additional variables were designed on the basis of results reported in McGrew, et al. (1994). A primary goal was to have a measure that could discriminate well-implemented ACT programs from other types of case management services and at the same time provide an accessible tool for training and self-evaluation within programs" (Zahrt et al., 1999).

Instrument length. 26-items rated on a 5 behaviorally-anchored response alternatives.

Reliability/validity sample. Teague (1998) examined the DACTS in 50 case management programs, representing four distinct types of service models: ACT, intensive case management provided by the Veterans Administration, outreach programs for people who were homeless and mentally ill, and traditional case management.

Data sources. "Informed" sources: team leader, case managers, objective records. Interview with team leader takes about an hour

Subscale structure. A factor analysis of the 26-item instrument yielded 8 factors, including 5 factors consisting of at least 3 items: Team & Intensity (5 items), Community Treatment (6 items), Engagement & Retention (4 items), Substance Abuse Treatment (3 items), and Specialist Staffing (3 items).

Reliability.

- Internal consistency (Cronbach's alpha). Total scale (.92), Team & Intensity (.87), Community Treatment (.87), Engagement & Retention (.77), Substance Abuse Treatment (.74), and Specialist Staffing (.77) in Teague et al. (1998) study; Winter and Calsyn (2000) and Zahrt et al. (Zahrt et al., 1999) had poorer internal consistency in more homogeneous samples.
- Stability of the scale. Low stability (Winter & Calsyn, 2000). Informal reports suggest change over time in direction of increased fidelity (Bond, Salyers, & Fekete, 1996).

Validity.

- Content validity. No formal evidence, but anecdotal reports suggest that it has good face validity

- Predictive validity. Zahrt (1999) found a correlation of.49 with reduction of in hospital use in a sample of 10 ACT programs.
- Known-groups validation. Teague et al. found that the DACTS discriminated across the four types of case management, consistent with predicted order of similarity to ACT. However, Johnsen et al. (1999) found that ACCESS programs (homeless outreach to people with SMI) ascribing to different labels (Strengths, ACT, etc.) don't actually look so different on the DACTS. Salyers et al. (1998) used the DACTS to discriminate between ACT and a "step-down" case management model intended to be a less intensive variation of ACT.

Other information.

- Comprehensiveness. The item coverage includes only a subset of the domains identified by experts as critical (McGrew & Bond, 1995).
- Floor/ceiling effect. Zahrt found no variation on 3 items.
- Agreement between data sources. Both Winter and Calsyn (2000) and Zahrt (1999) found mixed results for agreement between informants. Zahrt (1999) found good agreement on items relating to staffing and organizational factors, but poor agreement on items relating to service intensity.
- Use of scale since original study. It has been used as a monitoring tool in statewide ACT projects (Zahrt et al., 1999). It has also been used to monitor ACCESS projects over time (Winter & Calsyn, 2000) and is used by consultants in helping newly-developed ACT teams (Meisler, N., personal communication, June, 1999).

Recommendations. While some work is needed on improving individual items and there is a need to increasing the clinical dimensions of ACT, this instrument for the time being appears to be as good or better than any scales in use. Its use for discriminating between ACT programs and other types of case management appears to be adequate, and it appears useful as a management tool.

III. Vocational Program Scales

IPS Fidelity Scale

References. (Bond et al., 1997a)

Item generation. Items suggested by the model developers and followed suggestion in treatment manual (Becker & Drake, 1993).

Instrument length. 15 items rated on a 5 behaviorally-anchored response alternatives.

Reliability/validity sample. 27 sites including 9 IPS programs, 11 other SE programs, and 7 other vocational rehabilitation (VR) programs (Bond et al., 1997a). The IPS Fidelity Scale was also used in three subsequent studies (Bond et al., 1998b; Bond et al., 1999c; Vogler, 1998). With these additional data bases, the total sample was later expanded to 123 sites (Bond, 1999).

Data sources. Team leader interviews, site visits, objective records, and interviews with employment specialists. Interview takes about one hour

Subscale structure. 3 a priori scales, plus a total scale, in the original study. The a priori subscales were not statistically independent of each other: Staffing with Organization (r =.23, n.s.), Staffing with Service (r =.62, p <.001), and Organization with Service (r =.62, p <.001). In the 123-site sample, factor analysis yielded a 4-factor solution.

Reliability.

- Internal consistency (Cronbach's alpha). Total Scale (.92), Staffing (.72), Organization (.65), and Serce (.90).
- Agreement between interviewers. Very good. All but one item had interrater reliability of .80 or higher.

Validity.

- Known-groups validation. The IPS Fidelity Scale clearly discriminates between SE programs and non-SE programs. Both IPS and other SE programs are very different from other VR on all 15 items. The scale is only modestly useful for distinguishing between IPS and other SE programs, with differences showing up mainly in the area of integration of rehabilitation and treatment. Effect size (d) (Cohen, 1992) was used as the measure. Comparing IPS to Other VR, d was large: Total (2.48), Staffing (1.56), Organization (2.88), and Service (2.16). Similarly, comparing Other SE to Other VR, d was also large, for all but one subscale: Total (1.76), Staffing (1.67), Organization (0.02), and Service (1.64). Finally, comparing IPS to Other SE, d was large as well, for all but one subscale: Total (1.62), Staffing (-0.09), Organization (2.66), and Service (0.99).
- Predictive validity. Within relatively homogeneous samples there is little evidence that IPS Fidelity correlates with better employment rates (Bond et al., 1999c; Vogler, 1998). However, these results should be viewed cautiously because of difficulties defining and measuring employment outcomes.
- Concurrent validity. Correlation with SISTEM (described below): .67 (n = 24).

Other information.

- Comprehensiveness. Bond et al. (1997a) note that "Although serviceable in its current form, more refinement of existing items and development of additional items (especially for the Staffing and Organization subscales) will improve this instrument."
- Agreement between data sources. Vogler (1998) found generally high Pearson correlations on the Total Scale between supervisor and employment specialist (.75, n = 20) and between two employment specialists (.86, n = 9), but substantially lower for 6 follow-along specialists (r <.56). There was good agreement between supervisor and employment specialist on Organization (.70) and Service (.70), but not Staffing (.02).
- Floor/ceiling effects. Across all items, the scale values were used as follows: 5 (51%), 4 (21%), 3 (13%), 2 (3%), and 1 (13%) in the original study. In Vogler's (1998) study, scale value 5 was used 53% of time, and scale value 4 was used 23% of time.
- Use of scale since original study. Scale continues to be used in a variety of research projects for monitoring fidelity (Furlong-Norman, 1996).

Recommendations. This is a sensible tool for monitoring the development of IPS programs. It is briefer than would be desirable to maximize internal consistency, and its main use is the overall scale, although information on individual items may be helpful as feedback.

Quality of Supported Employment Implementation Scale (QSEIS)

References. (Bond et al., 1999c).

Item generation. Initial pool of items was based on the IPS Fidelity Scale. Further items were suggested by national advisory board of individuals chosen for their knowledge and expertise in SE. We also consulted descriptions of best practices (Ford, 1995; Hoff, 1997; MacDonald & Roberts, 1998; Marrone, 1996; Matrix, 1992).

Instrument length. 33-item interviewer-rated checklist obtained during a semi-structured interview. Items rated on a 5 behaviorally-anchored response alternatives.

Reliability/validity sample. 32 SE programs in New Jersey (n = 20) and Kansas (n = 12). 84% of the programs had been in existence for at least 3 years: KS (M = 5.07 years, SD = 3.18), and NJ (M = 6.71 years, SD = 3.08).

Data sources. Team leader interviews. Interview is up to 1.5 hours in length.

Subscale structure. 3 a priori scale (Staffing, Organization, and Services) did not hold up to internal consistency. Factor analysis did not yield usable results. 4 subscales were defined after the fact, based on conceptual grouping: Teamwork (3 items), Planning and Support (6 items), Rapid Job Search (3 items), Integration with Mental Health (5 items). The four subscales were statistically independent, with correlations between subscales ranging from -.26 to +.25.

Reliability.

- Internal consistency (Cronbach's alpha). Total scale (.51), Teamwork (.74), Planning and Support (.60), Rapid Job Search (.74), Integration with Mental Health (.62).
- Agreement between interviewers. Overall, pairs of interviewers attained 84% exact agreement at item level.

Validity.

- Content validity. Expert panel, as described above
- Known-groups validation. Not assessed in samples known to vary on adherence to SE. However, mean overall implementation was similar in both states, with somewhat different patterns, with NJ rating higher on Planning and Support, and KS rating higher on Integration of Mental Health and Rapid Job Search.
- Concurrent validity. QSEIS Total correlated with IPS Fidelity Scale (r=.36, n = 22). Corresponding a priori subscales correlated better: Staffing (r =.50), Organization (r =.59), Services (r =.37).
- Predictive validity. The QSEIS total scale and the 4 subscales were correlated with 9 indicators of employment outcomes, obtained from a retrospective survey completed by program directors in 24 of the programs. The total QSEIS score was not significantly correlated with any of the outcome measures. Planning and Support correlated positively with job tenure (r =.62), but was not related to annual VR

closure rate (r = -.15). Conversely, Rapid Job Search was negatively correlated with job tenure (r = -.56), while positively correlated with annual VR closure rate (r =.46).

Other information.

- Floor/ceiling effects. Possible ceiling effects. Mean ratings exceeded 4.0 on a 5-point scale, for 18 of 33 items.
- Agreement between data sources. Relatively poor agreement between ratings made by state administrator completing checklist without an interview and interviewer ratings.
- Use of scale since original study. Data collection continues.

Recommendations. More psychometric work is needed before we would wholeheartedly recommend this scale. Need to contrast the scale in sample of programs not using SE. More work is needed at the basic level of consensus on core ingredients of supported employment.

Scale for the Indiana Supported Employment Model (SISTEM)

References. (Vogler, 1998)

Item generation. Developed specifically for use in Indiana after consulting staff at the Supported Employment Consultation and Training Center in Anderson, IN and examining guidelines for use in their training. The items borrow from and are an expansion of the IPS Fidelity Scale.

Instrument length. 35-item interviewer-rated checklist obtained during a semi-structured interview. Items rated on a 5 behaviorally-anchored response alternatives.

Reliability/validity sample. 24 sites, 18 in Indiana and 6 in Minnesota; 19 had been in existing for at least a year, and 5 for less than a year. The 5 sites that had less than 1 year in existence did not differ from the 19 established sites.

Data sources. Interviews with 22 supervisors, 31 employment specialists, 6 follow-along specialists, and one job developer at the 24 sites; 28 completed during site visits and 32 by telephone. Telephone interview included 2 or more interviewers in all but one case. Interview takes about 1 hour. Agency records were consulted during site visits conducted at 13 sites.

Subscale structure. The original instrument was constructed around 3 a priori dimensions (Human Resources, Organization, and Services). The internal consistency of these subscales was low (.33 -.62). Factor analyses did not yield interpretable findings. A revised, 35-item scale consisting of 4 subscales (Human Resources, Organization, VR, and Services) was constructed on conceptual grounds.

Reliability.

- Internal consistency (Cronbach's alpha). Total Scale (.59), Human Resources (.48), Organization (56), VR (.05), and Services (.60)
- Agreement between interviewers. Interrater reliability (Total Scale Score intraclass correlation) =.97

Validity.

- Content validity. Reviewed by staff for Indiana supported employment technical assistance center.

- Known-groups validation. Comparisons between Indiana (n = 18) and Minnesota (n = 6) sites yielded one significant difference on 4 subscales and total scale: Indiana averaged higher ratings on Services. No differences between 13 rural and 11 urban sites.
- Concurrent validity. Correlation with IPS Fidelity Scale: .67 (n = 24).
- Predictive validity. Correlations between SISTEM total scale with 9 outcome measures (5 related to objective employment outcomes, 2 satisfaction scales, and 2 cost measures) yielded correlations ranging from .04 to .26. Correlations with SISTEM subscales were likewise generally small. Employment data were obtained from both agency records and VR reports. Satisfaction data were collected by the state department of mental health.

Other information.

- Comprehensiveness. An extension of IPS, possibly idiosyncratic to Indiana
- Agreement between data sources. For these analyses interviews were conducted with a supervisor and at least one employment specialist at 20 sites, with 9 additional sites in which there was a second employment specialist. There were 6 sites at which a follow-up specialist was interviewed. Pearson correlations between the supervisor and employment specialist at each site ranged from .74 to .91 for the Human Resources and Organization subscales. However, on the Services subscale, supervisor and employment specialist correlated .38 with the first employment specialist and .47 with the second employment specialist, while the two employment specialists correlated .71. Supervisor and employment specialists correlated .64 and the two employment specialists correlated .90 on the Total Score. Correlations between the follow-up specialist and both the supervisor and employment specialist were generally lower (.16 and .44, respectively, for the Total Score).
- Missing data analysis. 6.0% of all items were recorded as missing, including 2.9% wherein the respondent said they "did not know" and 3.1% where the interviewer rated the item as "does not fit."
- Floor/ceiling effects. Ratings positively skewed: 5 (58%), 4 (10%), 1-3 (32%). (Note: Percentages adjusted to ignore missing data.)

Recommendations. Study provides important and unique methodological data. Scale might be consulted as source of items if developing a scale. Some items may not be generalizable outside Indiana.

Diversified Placement Approach (DPA) Fidelity Scale (scale under development)

References. (Rollins et al., 2000)

Item generation. This scale is created to monitor the implementation of the DPA for a randomized controlled trial of two vocational programs. The DPA scale will be administered to the DPA program, as well as the other vocational program (IPS) in the study to detect and correct program drift during the study.

Items were generated through review of the existing literature on DPA, observation of DPA program, interviews and meeting with DPA staff and managers.

Instrument length. 20-items, rated on a 5-point, behaviorally-anchored scale.

Data source. Interview with team leaders and caseworkers and team meeting observation.

Subscale structure. 4 a priori dimensions: work readiness, staffing, integration of service, and array of options.

Progress. The scale and interview protocols have been piloted and revised, as of 2/2000.

Contact person for additional information. Angela L. Rollins or Gary Bond, Indiana University Purdue University, Indianapolis, Phone: (317) 274-6760, Fax: (317) 274-6756,

IV. Residential Program Scales

Ideologies of Care

References. (Heaney & Burke, 1995)

Item generation. Semi-structure interviews with staff during a pilot phase

Instrument length. 14-item checklist, with 7-point scale of importance

Reliability/validity sample. 192 house managers and 1653 direct care staff in 269 group homes for people with developmental disabilities or mental illness in Michigan.

Data sources. Residential staff

Subscale structure. Factor analysis yielded 4 factors: Normalization, Family Orientation, Protection of rights, and Business orientation. Two subscales : Normalization (5 items) and Family Orientation (3 items) were adapted from the factor analysis.

Reliability. Internal consistency (Cronbach's alpha). Normalization (.79) and Family Orientation (.80)

Validity. Not reported.

Other information. Direct care staff scored higher than house managers on Normalization and the opposite was true for Family Orientation

Recommendations. Because this instrument was piloted in a broader population than just people with SMI, it may not be appropriate for use in specifically group homes for people with SMI.

- Fidelity Assessment for the Center for Mental Health Services Housing Initiative (scale under development)

References. Lassiter (personal communication, December 3, 1999)

Goal. To understand the fidelity of housing programs or options to the specific structure and operation of various housing approaches (e.g., supported housing, supervised apartments, group homes).

Dimensions. The scale includes 12 dimensions: Housing choice, Separation of housing and services, Housing affordability, Integration, Rights of tenure, Service choice, Service individualization, Community-based service availability, Quality, Structure, Privacy, and Several descriptive dimensions, such as location and proximity to amenities in the area.

Progress. Currently the program manager and key staff informant instruments are under development, and are expected to be in the field by February 2000. The resident instrument is currently being used in the field,

as part of the 12-month interview for residents in study housing or as an exit interview for those who have left study housing.

Contact person for additional information. Debra Rog, Ph.D., Vanderbilt Institute for Public Studies, Phone: (202) 234-1190, Fax: (202) 234-1185

V. Drop-In Center Scales

See Table of Contents for cross-referenced scales

Member Involvement Scale

References. Mowbray, Robinson, and Holter (1999)

Item generation. Items were generated from meeting with Justice in Mental Health Organization staff and consumers in other drop-in centers at two statewide meetings

Instrument length. 10 open-ended items. Responses to each item were re-coded into 4 categories according to degree of consumer involvement (decisions being made by 1=Members or members with the board, 2=Board or board and director, 3=Executive Director and/or other staff, 4=Others outside the center)

Reliability/validity sample. 33 drop-in centers in Michigan

Data sources. 1-hour phone interview with director of the center.

Subscale structure. None.

Reliability. Internal consistency (Cronbach's alpha). .82

Validity. MIS differentiated between consumer-run and consumer-involved drop-in centers.

Recommendations. Though there is some evidence for known-grown validation, it requires more psychometric work. Authors recognized the need for a more comprehensive investigation including lager samples in different states, on-site observation and survey of consumers.

VI. Clubhouse Scales

Clubhouse Fidelity Index

References. Lucca (2000)

Item generation. The items were based on the theoretical and empirical literature (Beard et al., 1982; Horne & Otto, 1982; Macias, Jackson, Schroeder, & Wang, 1999; Mastbloom, 1992; Propst, 1992).

Instrument length. 15 dichotomous items

Reliability/validity sample. 22 psychiatric rehabilitation programs in Connecticut

Data sources. Brief interviews with program director or staff member. It takes 15-30 minutes.

Subscale structure. None

Reliability. Internal consistency (Cronbach's alpha). Total score (.75)

Validity.

- Known-groups validation. Discriminated among 3 a priori groups: programs certified by the International Center for Clubhouse Development, programs self-labeled as clubhouses but not certified, and other psychiatric rehabilitation programs.
- Concurrent validity. r =.57 with another scale of psychiatric rehabilitation values, also developed by the investigator.

Other information.

- Comprehensiveness. Focused narrowly on the vocational dimension; it does not capture all aspects of the clubhouse philosophy.
- Floor/ceiling effects. Range was from 7.9 (out of 18) for non-clubhouse to 16.5 for certified clubhouses. Lucca did not use two items that are not characteristic of clubhouses (medication administration and psychotherapy), because only one site had offered either of these.

Recommendations. Appears to be useful as a simple tool for classification for vocational services, but probably not useful for making fine discriminations, nor is it intended as a general-purpose tool.

VII. Skills Training Scales

Therapist Fidelity Evaluation for Skills Training Checklist

References. (Wallace et al., 1992)

Item generation. Developed from skills training manuals developed by Liberman et al.

Instrument length. 170 items in checklist format (60 of which are optional)

Reliability/validity sample. Seven treatment facilities (one long-term rehabilitation program, five residential care programs, one day treatment program) implemented three modules: medication management, recreation, and grooming

Data sources. Checklist completed while observing skills training groups by research assistants. Simple checklist requires observation of group either in person or on videotape.

Subscale structure. Not assessed.

Reliability.

- Internal consistency (Cronbach's alpha). Medication module (.87), Grooming module (.64), Recreation module (.73)
- Agreement between interviewers. Kappa was determined by calculating the agreement between the two sets of ratings of the checklist. Kappa =.88.

Validity. Not reported.

Other information. A similar version of the instrument is in use at the University of Chicago Center for Psychiatric Rehabilitation for self-monitoring purposes.

Recommendations. Can be useful in helping to delineate to group leaders the important components of skills training.

VIII. Family Psychoeducation Scales

☞ *Therapist Fidelity and Competency Scale (TFCS)*

References. (Mueser et al., 1994)

Item generation. For fidelity subscales, investigators generated items describing unique therapist behaviors to each model (Behavioral Family Therapy or Educational Family Therapy). Therapist fidelity was defined as "demonstrating at least five out of nine behaviors for one model, while demonstrating no more than one behavior that is unique to the other model" (p.104).

Instrument length. 9 dichotomous items for each of fidelity subscale and 4 items for each of competency subscales with 5-point Likert scales.

Reliability/validity sample. 8 families who received treatment at 3 hospitals

Data sources. Family therapy sessions over a year (mean 26.75) were audio taped. Number of items on subscales that were presented in each session was summed as an overall fidelity score. Of the audiotapes of family sessions, 10 sessions of each model were randomly selected and rated.

Subscale structure. The scale consists of 6 subscales: 2 therapist fidelity scales (Behavioral Family Therapy and Educational Family Therapy) and 4 therapist competency scales (Structure, Relationship, Difficulties, and Global)

Reliability.

- Internal consistency (Cronbach's alpha). Behavioral Family Therapy-Fidelity (.93), Educational Family Therapy-Fidelity (.87), Competency scales: Structure (.82), Relationship (.74), Difficulties (.69), Global (.86)
- Agreement between interviewers. Correlation between raters were .93 (BFT-Fidelity) and .87 (EFT-Fidelity), Structure (.82), Relationship (.74), Difficulties (.69), and Global (.86)

Validity.

- The TFCS discriminates between behavioral family therapy and educational family therapy though it did not find significant differences in general therapist competency skills between two therapy models.

Recommendations. This scale is based on pilot data and requires more psychometric work. Need to compare this with other models of family psychoeducation in experimental design.

IX. Supported Education Scales

None located as of 1/4/00

Supplemental Listing of Survey Instruments

These are instruments that have been used to describe programs, but are not scales per se, in the sense of giving scale values on specific dimensions.

Program Environment Scales

- Psychosocial Rehabilitation Program Survey (Lucca, 1998)
- Virginia Survey (McCall, 1994)

Case Management

- Case Management Practices Survey (Ellison et al., 1995)
- Intensive Case Management Survey (Schaedle, 1998)

Vocational Program Surveys

- Survey of Exemplary Supported Employment Programs (Gervey, Parrish, & Bond, 1995)
- Process Analysis of Supported Employment (Rogers, MacDonald-Wilson, Danley, Martin, & Anthony, 1997)

Residential Program Scales

- Consumer satisfaction with housing surveys (Tanzman, 1993)

Clubhouse Scales

- ICCD National Survey (Macias et al., 1999)
- Mastbloom survey (Mastbloom, 1992)
- Picone survey (Picone, Drake, Becker, Bond, & Anderson, 1998)

Statewide Fidelity Instruments

- Kansas Best Practice Fidelity Scales (Rapp, 1999)
- Illinois CONNECT98 Fidelity Scales (Bond et al., 1998b)
- Michigan Clubhouse Project (Onaga, 1999)
- Rhode Island Mobile Treatment Team Standards (Mazza, 1999)

References

Allness, D. J., & Knoedler, W. H. (1998). The PACT model of community-based treatment for persons with severe and persistent mental illness: A manual for PACT start-up. Arlington, VA: NAMI.

Anthony, W. A. (1980). The principles of psychiatric rehabilitation. Baltimore: University Park Press.

Anthony, W. A. (1994). Whither the "Boston University model"? Psychosocial Rehabilitation Journal, 17(4), 169-170.

Anthony, W. A., Buell, G. J., Sharratt, S., & Althoff, M. E. (1972). The efficacy of psychiatric rehabilitation. Psychological Bulletin, 78, 447-456.

Anthony, W. A., Cohen, M. R., & Cohen, B. F. (1984). Psychiatric rehabilitation. In J. A. Talbott (Ed.), The chronic mental patient: Five years later (pp. 137-157). Orlando, FL: Grune & Stratton.

Anthony, W. A., Cohen, M. R., & Farkas, M. D. (1982). A psychiatric rehabilitation program: Can I recognize one when I see one? Community Mental Health Journal, 18, 83-96.

Anthony, W. A., Cohen, M. R., & Vitalo, R. (1978). The measurement of rehabilitation outcome. Schizophrenia Bulletin, 4, 365-383.

Anthony, W. A., & Jansen, M. A. (1984). Predicting the vocational capacity of the chronically mentally ill: Research and implications. American Psychologist, 39, 537-544.

Anthony, W. A., & Liberman, R. P. (1992). Principles and practice of psychiatric rehabilitation. In R. P. Liberman (Ed.), Handbook of psychiatric rehabilitation (pp. 1-29). New York: Macmillan.

Austin, J. T. V., P. (1992). The criterion problem: 1917-1992. Journal of Applied Psychology, 77, 836-874.

Azrin, H. N., & Besalel, V. A. (1980). The job club counselor's manual: A behavioral approach to vocational counseling. Baltimore: University Park Press.

Azrin, N. H., & Philip, R. A. (1979). The job club method for the job handicapped: A comparative outcome study. Rehabilitation Counseling Bulletin, 23, 144-155.

Bachrach, L. L. (1988). The chronic patient: On exporting and importing model programs. Hospital and Community Psychiatry, 39, 1257-1258.

Bachrach, L. L. (1992). Psychosocial rehabilitation and psychiatry in the care of long-term patients. American Journal of Psychiatry, 149, 1455-1463.

Baronet, A., & Gerber, G. J. (1998). Psychiatric rehabilitation: Efficacy of four models. Clinical Psychology Review, 18, 189-228.

Bartko, J. J., & Carpenter, W. T. (1976). On the methods and theory of reliability. Journal of Nervous and Mental Disease, 163, 307-317.

Barton, R. (1997). Psychosocial rehabilitation services and community support systems: Outcomes research and policy implications. Springfield, IL: Office of Mental Health of the Illinois Department of Human Services.

Barton, R. (1999). Psychosocial rehabilitation services in community support systems: A review of outcomes and policy recommendations. Psychiatric Services, 50, 525-534.

Beard, J. H., Propst, R. N., & Malamud, T. J. (1982). The Fountain House model of rehabilitation. Psychosocial Rehabilitation Journal, 5(1), 47-53.

Beard, M. L. (1992). Social networks. Psychosocial Rehabilitation Journal, 16(2), 111-123.

Becker, D. R., & Drake, R. E. (1993). A working life: The Individual Placement and Support (IPS) Program. Concord, NH: New Hampshire-Dartmouth Psychiatric Research Center.

Becker, D. R., Drake, R. E., Farabaugh, A., & Bond, G. R. (1996). Job preferences of clients with severe psychiatric disorders participating in supported employment programs. Psychiatric Services, 47, 1223-1226.

Becker, D. R., Torrey, W. C., Toscano, R., Wyzik, P. F., & Fox, T. S. (1998a). Building recovery-oriented services: Lessons from implementing IPS in community mental health centers. Psychiatric Rehabilitation Journal, 22(1).

Becker, T., Holloway, F., McCrone, P., & Thornicroft, G. (1998b). Evolving service interventions in Nunhead and Norwood: PRiSM Psychosis Study 2. British Journal of Psychiatry, 173, 371-375.

Bond, G. R. (1991). Variations in an assertive outreach model. New Directions for Mental Health Services, 52, 65-80.

Bond, G. R. (1998). Principles of the Individual Placement and Support model: Empirical support. Psychiatric Rehabilitation Journal, 22(1), 11-23.

Bond, G. R. (1999, November 2). Critical elements of supported employment. Paper presented at the American Psychiatric Association Institute on Psychiatric Services, New Orleans, LA.

Bond, G. R., Becker, D. R., Drake, R. E., & Vogler, K. M. (1997a). A fidelity scale for the Individual Placement and Support model of supported employment. Rehabilitation Counseling Bulletin, 40, 265-284.

Bond, G. R., & colleagues. (in progress). Comparison of two employment models for clients with SMI : NIMH Grant # 1 R01 MH59987.

Bond, G. R., Dietzen, L. L., Vogler, K. M., Katuin, C. H., McGrew, J. H., & Miller, L. D. (1995). Toward a framework for evaluating costs and benefits of psychiatric rehabilitation: Three case examples. Journal of Vocational Rehabilitation, 5, 75-88.

Bond, G. R., Drake, R. E., & Becker, D. R. (1998a). The role of social functioning in vocational rehabilitation. In K. T. Mueser & N. Tarrier (Eds.), Handbook of social functioning in schizophrenia (pp. 372-390). Needham Heights, MA: Allyn & Bacon.

Bond, G. R., Drake, R. E., Becker, D. R., & Mueser, K. T. (1999a). Effectiveness of psychiatric rehabilitation approaches for employment of people with severe mental illness. Journal of Disability Policy Studies, 10(1), 18-52.

Bond, G. R., Drake, R. E., Mueser, K. T., & Becker, D. R. (1997b). An update on supported employment for people with severe mental illness. Psychiatric Services, 48, 335-346.

Bond, G. R., Evans, L., Kim, H., & Goodman, C. (1999b). Evaluation of Connect98. Phase II: Site visits by network staff (Final report to Illinois Office of Mental Health). Indianapolis, IN: Indiana University Purdue University Indianapolis.

Bond, G. R., Evans, L., Salyers, M. P., Williams, J., & Kim, H. K. (2000). Measurement of fidelity in psychiatric rehabilitation. Mental Health Services Research, 2, 75-87.

Bond, G. R., Evans, L. J., & Resnick, S. G. (1998b). Evaluation of Connect 98 (Final Report to Illinois Office of Mental Health). Indianapolis, IN: Indiana University-Purdue University Indianapolis.

Bond, G. R., Picone, J., Mauer, B., Fishbein, S., & Stout, R. (1999c). The Quality of Supported Employment Implementation Scale. In G. Revell, K. J. Inge, D. Mank, & P. Wehman (Eds.), The impact of supported employment for people with significant disabilities: Preliminary findings from the National Supported Employment Consortium (pp. 73-88). Richmond, VA: Virginia Commonwealth University.

Bond, G. R., Picone, J., & Mauer, E. (1998c). The Quality of Supported Employment Implementation Scale.

Bond, G. R., & Resnick, S. G. (2000). Psychiatric rehabilitation. In R. G. Frank & T. Elliott (Eds.), Handbook of rehabilitation psychology (pp. 235-258). Washington, DC: American Psychological Association.

Bond, G. R., Salyers, M. P., & Fekete, D. M. (1996). Illinois ACT Project: Final report. Indianapolis, IN: Indiana University Purdue University Indianapolis.

Bond, G. R., Vogler, K., & Irmscher, S. (1994). North Dakota Clubhouse Evaluation: Final Report. Indianapolis, IN: Indian University Purdue University Indianapolis.

Bond, G. R., Witheridge, T. F., Dincin, J., Wasmer, D., Webb, J., & DeGraaf-Kaser, R. (1990). Assertive community treatment for frequent users of psychiatric hospitals in a large city: A controlled study. American Journal of Community Psychology, 18, 865-891.

Boyer, S. L., & Bond, G. R. (1999). Does assertive community treatment reduce burnout? A comparison with traditional case management. Mental Health Services Research, 1, 31-45.

Brekke, J. S. (1987). The model-guided method for monitoring program implementation. Evaluation Review, 11(3), 281-299.

Brekke, J. S. (1988). What do we really know about community support programs? Strategies for better monitoring. Hospital and Community Psychiatry, 39, 946-952.

Brekke, J. S., & Aisley, R. A. (1990). The client interaction scale. Evaluation and the Health Professions, 13, 215-226.

Brekke, J. S., & Test, M. A. (1987). An empirical analysis of services delivered in a model community support program. Psychosocial Rehabilitation Journal, 10(4), 51-61.

Brekke, J. S., & Test, M. A. (1992). A model for measuring the implementation of community support programs: Results from three sites. Community Mental Health Journal, 28, 227-247.

Brekke, J. S., & Wolkon, G. H. (1988). Monitoring program implementation in community mental health settings. Evaluation & the Health Professions, 11, 425-440.

Brook, R. H. (1989). Practice guidelines and practicing medicine: Are they compatible? Journal of American Medical Association, 262, 3027-3030.

Burns, T., Creed, F., Fahy, T., Thompson, S., Tyrer, P., & White, I. (1999). Intensive versus standard case management for severe psychotic illness: a randomised trial. UK 700 Group. Lancet, 353, 2185-2189.

Burt, M., Duke, A. E., & Hargreaves, W. A. (1998). The Program Environment Scale: Assessing client perceptions of community-based programs for the severely mentally ill. American Journal of Community Psychology, 26, 853-879.

Burt, M., & Hargreaves, W. A. (1997, November). Developing the Program Environment Scale: A tool to assess the perceptions of severely mentally ill program clients in community-based programs. Paper presented at the American Public Health Association, Indianapolis, IN.

Calsyn, R. J. (2000). A checklist for critiquing treatment fidelity studies. Mental Health Services Research, 2, 107-113.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56, 81-105.

CARF. (2000). 2000 Behavioral Standards Manual. Tucson, AZ: CARF, the Rehabilitation Accreditation Commission.

Carkhuff, R. R. (1969). Helping and human relations. (Vol. 1 & 2). New York: Holt, Rinehart, & Winston.

Carling, P. J. (1993). Housing and supports for persons with mental illness: Emerging approaches to research and practice. Hospital and Community Psychiatry, 44, 439-449.

Cascio, W. F. (1998). Applied psychology in human resource management. Upper Saddle River: Prentice Hall.

Chandler, D., Levin, S., & Barry, P. (1999). The menu approach to employment services: Philosophy and five-year outcomes. Psychiatric Rehabilitation Journal, 23(1), 24-33.

Clark, R. E., Teague, G. B., Ricketts, S. K., Bush, P. W., Keller, A. M., Zubkoff, M., & Drake, R. E. (1994). Measuring resource use in economic evaluations: Determining the social costs of mental illness. Journal of Mental Health Administration, 21, 32-41.

Cnaan, R. A., Blankertz, L., Messinger, K. W., & Gardner, J. R. (1988). Psychosocial rehabilitation: Toward a definition. Psychosocial Rehabilitation Journal, 11(4), 61-77.

Cnaan, R. A., Blankertz, L., Messinger, K. W., & Gardner, J. R. (1990). Experts' assessment of psychosocial rehabilitation principles. Psychosocial Rehabilitation Journal, 13(3), 59-73.

Cochrane, J., Durbin, J., & Goering, P. (1997). Best practices in mental health reform: Discussion paper. Toronto, ON: Clarke Institute of Psychiatry.

Cohen, J. (1992). A power primer. Psychological Bulletin, 112, 155-159.

Conway, J. M., Jako, R. A., & Goodman, D. F. (1995). A meta-analysis of interrater and internal consistency reliability of selection interviews. Journal of Applied Psychology, 80, 565-579.

Corrigan, P. W., Reedy, P., Thadani, D., & Ganet, M. (1995). Correlates of participation and completion in a job club for clients with psychiatric disability. Rehabilitation Counseling Bulletin, 39, 42-53.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley.

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. Psychological Bulletin, 81, 95-106.

Deci, P. A., Santos, A. B., Hiott, D. W., Schoenwald, S., & Dias, J. K. (1995). Dissemination of assertive community treatment teams. Psychiatric Services, 46, 676-678.

Dilk, M. N., & Bond, G. R. (1996). Meta-analytic evaluation of skills training research for individuals with severe mental illness. Journal of Consulting and Clinical Psychology, 64, 1337-1346.

Dincin, J. (1975). Psychiatric rehabilitation. Schizophrenia Bulletin, 1, 131-147.

Dincin, J. (1988). A crucial dimension, Switzer Monograph (pp. 25-47). Alexandria, VA: National Rehabilitation Association.

Dincin, J. (1995). A pragmatic approach to psychiatric rehabilitation: Lessons from Chicago's Thresholds program. New Directions for Mental Health Services, 69(Whole Issue).

Dion, G. L., & Anthony, W. A. (1987). Research in psychiatric rehabilitation: A review of experimental and quasi–experimental studies. Rehabilitation Counseling Bulletin, 30, 177–182.

Drake, R. E., Becker, D. R., Biesanz, J. C., Torrey, W. C., McHugo, G. J., & Wyzik, P. F. (1994). Rehabilitation day treatment vs. supported employment: I. Vocational outcomes. Community Mental Health Journal, 30, 519-532.

Drake, R. E., McHugo, G. J., Becker, D. R., Anthony, W. A., & Clark, R. E. (1996). The New Hampshire study of supported employment for people with severe mental illness: Vocational outcomes. Journal of Consulting and Clinical Psychology, 64, 391-399.

Drebing, C., & Van Ormer, A. (1999). Compensated Work Therapy: Blue ribbon panel on program variables related to outcome. Bedford, MA: New England Mental Illness Research Education and Clinical Center.

Eddy, D. M. (1990a). Clinical decision-making: From theory to practice: Practice policies: Where do they come from? Journal of American Medical Association, 263, 1265-1275.

Eddy, D. M. (1990b). Clinical decision-making: From theory to practice: The challenge. Journal of American Medical Association, 263, 287-290.

Ellison, M. L., Rogers, E. S., Sciarappa, K., Cohen, M., & Forbess, R. (1995). Characteristics of mental health case management: Results of a national survey. Journal of Mental Health Administration, 22, 101-112.

Estroff, S. (1981). Making it crazy. Berkeley, CA: University of California Press.

Evans, L. J., Resnick, S. G., & Bond, G. R. (1998). The Psychiatric Rehabilitation Environment Scale..

Eysenck, H. (1952). The effects of psychotherapy, an evaluation. Journal of Consulting Psychology, 16, 319-324.

Fairweather, G. W. (1980). The Fairweather lodge: A twenty-five year retrospective. New Directions for Mental Health Services, 7(Whole Issue).

Fairweather, G. W., Sanders, D., Cressler, D., & Maynard, H. (1969). Community life for the mentally ill: An alternative to hospitalization. Chicago, IL: Aldine.

Farkas, M. D., & Anthony, W. A. (Eds.). (1989). Psychiatric rehabilitation programs: Putting theory into practice. Baltimore: Johns Hopkins University Press.

Farkas, M. D., Cohen, M. R., & Nemec, P. B. (1988). Psychiatric rehabilitation programs: Putting concepts into practice? Community Mental Health Journal, 24, 7-21.

Fergus, E. O., & Balzell, A. (1990). A national directory of Fairweather programs. East Lansing, MI: Michigan State University Fairweather Lodge Project.

Fishbein, S. M. (1988). Partial care as a vehicle for rehabilitation of individuals with severe psychiatric disability. Rehabilitation Psychology, 33, 57-64.

Fiske, D. W. (1971). Measuring the concepts of personality. Chicago: Aldine.

Flexer, R. W., & Solomon, P. L. (Eds.). (1993). Psychiatric rehabilitation in practice. Boston: Andover Medical Publishers.

Ford, L. H. (1995). Providing employment support for people with long-term mental illness. Baltimore: Paul H. Brookes.

Fowler, F. J. (1988). Survey research methods. Newbury Park: Sage.

Fowler, F. J. (1995). Improving survey questions: Design and Evaluation. Thousand Oaks: Sage.

Frances, A., Docherty, J. P., & Kahn, D. A. (1996). The expert consensus guideline series: Treatment of schizophrenia. Journal of Clinical Psychiatry, 57(Supplement 12B), 1-58.

Furlong-Norman, K. (1996). CMHS Employment Intervention Demonstration Program. Community Support Network News, 11(3), (Whole Issue).

Gervey, R., Parrish, A., & Bond, G. R. (1995). Survey of exemplary supported employment programs for persons with psychiatric disabilities. Journal of Vocational Rehabilitation, 5, 115-125.

Giesler, L. J., & Hodge, M. (1998). Case management in behavioral health care. International Journal of Mental Health, 27, 26-40.

Gold Award. (1999). The wellspring of the clubhouse model for social and vocational adjustment of persons with serious mental illness: Fountain House, New York City. Psychiatric Services, 50, 1473-1476.

Gorey, K. M., Leslie, D. R., Morris, T., Carruthers, W. V., John, L., & Chacko, J. (1998). Effectiveness of case management with severely and persistently mentally ill people. Community Mental Health Journal, 34, 241-250.

Guadagnoli, E., & Velicer, W. F. (1988). Relation of sample size to the stability of component patterns. Psychological Bulletin, 103, 265-275.

Halford, W., & Hayes, R. (1991). Psychosocial rehabilitation of chronic schizophrenic patients: Recent findings on social skills training and family psychoeducation. Clinical Psychology Review, 11, 23-44.

Hall, L. L., Edgar, E. R., & Flynn, L. M. (1997). Stand and deliver: Action call to a failing industry : National Alliance for the Mentally Ill.

Hargreaves, W. A., Shumway, M., Hu, T. W., & Cuffel, B. (1998). Cost-outcome methods for mental health. San Diego: Academic Press.

Hatfield, A. B. (1989). Serving the unserved in community rehabilitation programs. Psychosocial Rehabilitation Journal, 13(2), 71-82.

Heaney, C. A., & Burke, A. C. (1995). Ideologies of care in community residential services: What do caretakers believe? Community Mental Health Journal, 31, 449-462.

Heflinger, C. A. (1996). Implementing a system of care: Finding from the Fort Bragg Evaluation Project. Journal of Mental Health Administration, 23, 16-29.

Henggeler, S. W., Pickrel, S. G., & Brondino, M. J. (1999). Multisystemic treatment of substance-abusing and -dependent delinquents: Outcomes, treatment fidelity, and transportability. Mental Health Services Research, 1, 171-184.

Hinkin, T. R. (1995). A review of scale development practices in the study of organizations. Journal of Management, 21, 967-988.

Hoff, D. (1997). Quality employment services: Will you know it when you see it?. Boston, MA: Institute of Community Inclusion, Children's Hospital.

Hogarty, G., Anderson, C., Reiss, D., Kornblith, S., Greenwald, D., Javna, D., & Madonia, M. (1986). Family psychoeducation, social skills training, and maintenance chemotherapy in the aftercare treatment of schizophrenia: One-year effects of a controlled study on relapse and expressed emotion. Archives of General Psychiatry, 43, 633-642.

Hogarty, G. E. (1995). Schizophrenia and modern mental health services. Decade of the Brain, 6(1), 3-6.

Horne, R., & Otto, F. (1982). Adirondack House: The evolution of a psychosocial clubhouse. Psychosocial Rehabilitation Journal, 6(2), 2-14.

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

Hughes, R. (1999). The meaning of "evidence based" services in PSR. IAPSRS Connection, 2, 1, 10-12.

Hughes, R., & Clement, J. (1999, March 1). Time to end the model wars. IAPSRS Connection, 1, 1.

IAPSRS. (1997a). IAPSRS position paper: The single model trap. Columbia, MD: International Association of Psychosocial Rehabilitation Services.

IAPSRS. (1997b). Practice guidelines for the psychiatric rehabilitation of persons with severe and persistent mental illness in a managed care environment. Columbia, MD: International Association of Psychosocial Rehabilitation Services.

Jacobs, H. E., Kardashian, S., Kreinbring, R. K., Ponder, R., & Simpson, A. S. (1984). A skills-oriented model for facilitating employment among psychiatrically disabled persons. Rehabilitation Counseling Bulletin, 28, 87-96.

Jacobs, H. E., Wissusik, D., Collier, R., Stackman, D., & Burkeman, D. (1992). Correlations between psychiatric disabilities and vocational outcome. Hospital and Community Psychiatry, 43, 365-369.

James, L. R., Demaree, R. G., & Wolf, G. (1995). rwg: An assessment of within-group interrater agreement. Journal of Applied Psychology, 78, 306-309.

Jerrell, J. M., & Hargreaves, W. A. (1991). The operating philosophy of community programs (Working Paper 18). Berkeley, CA: Institute for Mental Health Services Research.

Johnsen, M., Samberg, L., Calsyn, R., Blasinsky, M., Landow, W., & Goldman, H. (1999). Case management models for persons who are homeless and mentally ill: The ACCESS Demonstration Project. Community Mental Health Journal, 35, 325-346.

Kerlinger, F. N. (1986). Foundations of behavioral research. (3 ed.). New York: Holt, Rinehart, & Winston.

Lam, D. (1991). Psychosocial family intervention in schizophrenia: A review of empirical studies. Psychological Medicine, 21, 423-441.

Latimer, E. (1999a). Conseil d'évaluation des technologies de la santé du Québec. Suivi intensif en équipe dans la communauté pour personnes atteintes de troubles mentaux graves. (CETS 99-1 RF). Montréal: CETS.

Latimer, E. (1999b). Economic impacts of assertive community treatment: A review of the literature. Canadian Journal of Psychiatry, 44, 443-454.

Leff, J., Kuipers, L., Berkowitz, R., & Sturgeon, D. (1985). A controlled trial of social intervention in the families of schizophrenic patients: Two year follow-up. British Journal of Psychiatry, 146, 594-600.

Lehman, A. F., Steinwachs, D. M., & PORT Co-Investigators. (1998). At issue: Translating research into practice: The Schizophrenia Patient Outcomes Research Team (PORT) treatment recommendations. Schizophrenia Bulletin, 24, 1-10.

Liberman, R. P. (1985). Social skills training for chronic mental patients. Hospital and Community Psychiatry, 36, 396-403.

Liberman, R. P. (Ed.). (1988). Psychiatric rehabilitation of chronic mental patients. Washington, DC: American Psychiatric Association.

Lipsey, M. W. (1990). Design sensitivity. Newbury Park, CA: Sage.

Lucca, A. M. (1998). Relationships among program components, social environment, and member functioning in psychosocial rehabilitation programs for seriously mentally ill individuals. Unpublished dissertation, University of Connecticut.

Lucca, A. M. (2000). A clubhouse fidelity index: Preliminary reliability and validity results. Mental Health Services Research, 2, 89-94.

MacDonald, R., & Roberts, M. (1998). Supported employment quality service indicators : Boggs Center -- UAP and NJ APSE (Available from Melissa Roberts, UMDNJ, 1776 Raritan Road, Scotch Plains, NJ 07076).

Macias, C., & Jackson, R. (1993). The Self-Report Inventory: An interview schedule for adults with serious mental illness (Unpublished questionnaire). New York: Fountain House.

Macias, C., Jackson, R., Schroeder, C., & Wang, Q. (1999). What is a clubhouse? Report on the ICCD 1996 survey of USA clubhouses. Community Mental Health Journal, 35, 181-190.

Macias, C., Kinney, R., & Rodican, C. (1995). Transitional employment: An evaluative description of Fountain House practice. Journal of Vocational Rehabilitation, 5, 151-158.

Marrone, J. (1996). Managing employment with care: Better practices in employment for a state/county/provincial department of mental health to fund, support and monitor. Boston, MA: Institute for Community Inclusion, Children's Hospital.

Marshall, M., & Creed, F. (2000). Assertive community treatment: Is it the future of community care in the UK? Social Psychiatry and Psychiatric Epidemiology, 12, 191-196.

Marshall, M., Lockwood, A., Green, R., & Gray, A. (1998). Case management for people with severe mental disorders (Cochrane Review), Cochrane Library (Updated quarterly ed.,). Oxford, England: Update Software.

Marty, D., Rapp, C. A., & Carlson, L. (under review). The experts speak: The critical ingredients of strengths case management.

Mastbloom, J. (1992). Forty clubhouses: Models and practices. Psychosocial Rehabilitation Journal, 16(2), 9-23.

Matrix. (1992). Quality indicators for supported employment programs servicing persons with cognitive, psychiatric, and physical disabilities (Mid Atlantic Regional Information Exchange). Philadelphia, PA: Matrix Research Institute.

Mazza, G. (1999). Personal communication.

McCall, B. (1994). Survey of psychosocial rehabilitation programs. Research Notes. Richmond, VA: VA Department of Mental Health, Mental Retardation, & Substance Abuse Services.

McCarthy, D., Thompson, D., & Olson, S. (1998). Planning a statewide project to convert day treatment to supported employment. Psychiatric Rehabilitation Journal, 22(1), 30-33.

McDaniel, M. A., Whetzel, D. L., Shcmidt, F. L., & Maurer, S. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. Journal of Applied Psychology, 79, 599-616.

McDonnell, J., Nofs, D., Hardman, M., & Chambless, C. (1989). An analysis of the procedural components of supported employment programs associated with employment outcomes. Journal of Applied Behavior Analysis, 22, 417-428.

McEvoy, J. P., Scheifler, P. L., & Frances, A. (1999). The expert consensus guideline series: treatment of schizophrenia 1999. Journal of Clinical Psychiatry, 11(Supplement), 1-80.

McFarlane, W., Lukens, E., Link, B., Dushay, R., Deakins, S., Newmark, M., Dunne, E., Horen, B., & Toran, J. (1995). Multiple-family groups and psychoeducation in the treatment of schizophrenia. Archives of General Psychiatry, 52, 679-687.

McGrew, J. H., & Bond, G. R. (1995). Critical ingredients of assertive community treatment: Judgments of the experts. Journal of Mental Health Administration, 22, 113-125.

McGrew, J. H., Bond, G. R., Dietzen, L. L., & Salyers, M. P. (1994). Measuring the fidelity of implementation of a mental health program model. Journal of Consulting and Clinical Psychology, 62, 670-678.

McGrew, J. H., Wilson, R., & Bond, G. R. (1996). Client perspectives on helpful ingredients of assertive community treatment. Psychiatric Rehabilitation Journal, 19(3), 13-21.

McHugo, G. J., Drake, R. E., Teague, G. B., & Xie, H. (1999). The relationship between model fidelity and client outcomes in the New Hampshire Dual Disorders Study. Psychiatric Services, 50, 818-824.

McHugo, G. J., Hargreaves, W. A., Drake, R. E., Clark, R. E., Xie, H., Bond, G. R., & Burns, B. J. (1998). Methodological issues in assertive community treatment studies. American Journal of Orthopsychiatry, 68, 246-260.

Meehl, P. E. (1954). Clinical versus statistical prediction: A theoretical analysis and review of the evidence. Minneapolis: University of Minnesota Press.

Meisler, N. (1997). Assertive community treatment initiatives: Results from a survey of selected state mental health authorities. Community Support Network News, 11(4), 3-5.

Moncher, F. J., & Prinz, R. J. (1991). Treatment fidelity in outcome studies. Clinical Psychology Review, 11, 247-266.

Moos, R. H. (1974a). Community Oriented Program Environment Scale. Palo Alto, CA: Consulting Psychologist Press.

Moos, R. H. (1974b). Evaluation of treatment environments: A social ecological approach. New York: Wiley-Interscience.

Morse, G. A., Calsyn, R. J., Allen, G., Tempelhoff, B., & Smith, R. (1992). Experimental comparison of the effects of three treatment programs for homeless mentally ill people. Hospital and Community Psychiatry, 43, 1005-1010.

Mowbray. (1999). Drop-in center fidelity scale.

Mowbray, C. T., Chamberlain, P., Jennings, M., & Reed, C. (1988). Consumer-run mental health services: Results from five demonstration projects. Community Mental Health Journal, 2, 151-156.

Mowbray, C. T., Moxley, D. P., Jasper, C. A., & Howell, L. L. (Eds.). (1997). Consumers as providers in psychiatric rehabilitation. Columbia, MD: International Association of Psychosocial Rehabilitation Services.

Mowbray, C. T., Plum, T. B., & Masterton, T. (1998). Harbinger II: Deployment and evolution of assertive community treatment in Michigan. Administration and Policy in Mental Health, 25, 125-139.

Mowbray, C. T., & Tan, C. (1992). Evaluation of an innovative consumer-run service model: the drop-in center`. Innovations and Research, 1(2), 19-24.

Moxley, D. P. (1993). Clubhouse standards as a program development tool. Psychosocial Rehabilitation Journal, 17(2), 177-183.

Moxley, D. P., Mowbray, C. T., & Brown, K. S. (1993). Supported education. In R. W. Flexer & P. L. Solomon (Eds.), Psychiatric rehabilitation in practice (pp. 137-153). Boston: Andover Medical Publishers.

Mueser, K. T., Bond, G. R., Drake, R. E., & Resnick, S. G. (1998). Models of community care for severe mental illness: A review of research on case management. Schizophrenia Bulletin, 24, 37-74.

Mueser, K. T., Drake, R. E., & Bond, G. R. (1997). Recent advances in psychiatric rehabilitation for patients with severe mental illness. Harvard Review of Psychiatry, 5, 123-137.

Mueser, K. T., Drake, R. E., Clark, R. E., McHugo, G. J., Mercer-McFadden, C., & Ackerson, T. H. (1995). Evaluating substance abuse in persons with severe mental illness (HSRI Toolkit). Cambridge: the Evaluation Center@HSRI.

Mueser, K. T., Gingerich, S. L., & Rosenthal, C. K. (1994). Educational family therapy for schizophrenia: A new treatment model for clinical service and research. Schizophrenia Research, 13, 99-108.

Mueser, K. T., & Glynn, S. M. (1995). Families as members of the treatment team, Behavioral family therapy for psychiatric disorder (pp. 1-29). Needham Heights, MA: Allyn & Bacon.

Mueser, K. T., & Glynn, S. M. (1999). Behavioral family therapy for psychiatric disorders. (Second ed.). Oakland, CA: New Harbinger.

Noble, J. H. (1991). The benefits and costs of supported employment for people with mental illness and with traumatic brain injury in New York State (C-0023180). Buffalo, NY: Research Foundation of the State University of New York.

Noble, J. H., Honberg, R. S., Hall, L. L., & Flynn, L. M. (1997). A legacy of failure: The inability of the federal-state vocational rehabilitation system to serve people with severe mental illness. Arlington, VA: National Alliance for the Mentally Ill.

Nunnally, J. C., & Bernstein, I. H. (1994). Psychometric theory. New York: McGraw-Hill.

OMH. (1997). CONNECT98: A model for comprehensive psychosocial rehabilitation services. Springfield, IL: Office of Mental Health (OMH) of the Illinois Department of Human Services.

Onaga, E. (1999). Study in progress.

Orlinsky, D. E., Grawe, K., & Parks, B. K. (1994). Process and outcome in psychotherapy -- Noch einmal. In A. E. Bergin & S. L. Garfield (Eds.), Handbook of psychotherapy and behavior change (4th ed., pp. 270-376). New York: John Wiley.

Paul, G. L., & Lentz, R. J. (1977). Psychosocial treatment of chronic mental patients: Milieu versus social learning programs. Cambridge, MA: Harvard University Press.

Pedhazur, E. J., & Schmelkin, L. P. (1991). Measurement, design, and analysis: An integrated approach. Hillsdale: Lawrence Erlbaum.

Penn, D. L., & Mueser, K. T. (1996). Research update on the psychosocial treatment of schizophrenia. American Journal of Psychiatry, 15, 607-617.

Picone, J., Drake, R. E., Becker, D., Bond, G. R., & Anderson, L. (1998). A survey of clubhouse programs (unpublished). Indianapolis, IN: IUPUI.

Pratt, C. W., Gill, K. J., Barrett, N. M., & Roberts, M. M. (1999). Psychiatric rehabilitation. San Diego, CA: Academic Press.

Propst, R. (1992). Standards for clubhouse programs: Why and how they were developed. Psychosocial Rehabilitation Journal, 16(2), 25-30.

Randolph, E., Eth, S., Glynn, S., Paz, G., Leong, G., Shaner, A., Strachan, A., Van Vort, W., Escobar, J., & Liberman, R. (1994). Behavioral family management in schizophrenia: Outcome from a clinic-based intervention. British Journal of Psychiatry, 144, 501-506.

Rapp, C. A. (1998). The active ingredients of effective case management: A research synthesis. Community Mental Health Journal, 34, 363-380.

Rapp, C. A. (1999). Best Practice Fidelity Tools. Lawrence, KS: University of Kansas School of Social Welfare.

Rogers, C. R. (1951). Client-centered therapy. Boston: Houghton Mifflin.

Rogers, C. R. (1957). The necessary and sufficient conditions of therapeutic personality change. Journal of Consulting Psychology, 21, 95-103.

Rogers, E. S., MacDonald-Wilson, K., Danley, K., Martin, R., & Anthony, W. A. (1997). A process analysis of supported employment services for persons with serious psychiatric disability: Implications for program design. Journal of Vocational Rehabilitation, 8.

Rollins, A., Bond, G. R., Salyers, M. P., Resnick, S., Dincin, J., McCoy, M., Kinley, T., Shimon, S., Marcelle, K., Fraser, G., & Forman, J. (2000). Diversified Placement Approach Fidelity Scale. Chicago: Thresholds.

Rosenfield, S., & Neese-Todd, S. (1993). Elements of a psychosocial rehabilitation program associated with a satisfying quality of life. Hospital and Community Psychiatry, 44, 76-78.

Rosenheck, R., Neale, M., Leaf, P., Milstein, R., & Frisman, L. (1995). Multisite experimental cost study of intensive psychiatric community care. Schizophrenia Bulletin, 21, 129-140.

Rutman, I. D. (1993). And now, the envelope please... Psychosocial Rehabilitation Journal, 16(3), 1-3.

Ryan, C. S., Sherman, P. S., & Bogart, L. M. (1997). Patterns of services and consumer outcome in an intensive case management program. Journal of Consulting and Clinical Psychology, 65(3), 485-493.

Ryan, E. R., Bell, M. D., & Metcalf, J. C. (1982). The development of a rehabilitation psychology program for persons with schizophrenia: Changes in the treatment environment. Rehabilitation Psychology, 27, 67-85.

Salyers, M. P., Masterton, T. W., Fekete, D. M., Picone, J. J., & Bond, G. R. (1998). Transferring clients from intensive case management: Impact on client functioning. American Journal of Orthopsychiatry, 68, 233-245.

Schaedle, R. (1998). Critical ingredients of intensive case management: Judgments of experts, program managers, and case managers. Unpublished dissertation in progress, New York.

Schaedle, R., & Epstein, I. (2000). Specifying intensive case management: A multiple stakeholder approach. Mental Health Services Research (2), 95-105.

Scott, A. G., & Sechrest, L. (1989). Strength of theory and theory of strength. Evaluation and Program Planning, 12, 329-336.

Sechrest, L., West, R. G., Phillips, M. A., Redner, R., & Yeaton, W. (1979). Some neglected problems in evaluation and research: Strength and integrity of treatments. In L. Sechrest, S. G. West, M. A. Phillips, R. Redner, & W. Yeaton (Eds.) (Vol. 4, pp. 15-35). Beverly Hills, CA: Sage.

Sherman, P. S., & Porter, R. (1991). Mental health consumers as case management aides. Hospital and Community Psychiatry, 42, 494-498.

Shern, D. L., Trochim, W. M., & LaComb, C. A. (1995). The use of concept mapping for assessing fidelity of model transfer: An example from psychiatric rehabilitation. Evaluation and Program Planning, 18, 143.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin, 86, 399-441.

Silverman, S. H., Blank, M. B., & Taylor, L. C. (1997). On our own: Preliminary findings from a consumer-run service model. Psychiatric Rehabilitation Journal, 21(2), 151-156.

Solomon, P. (1992). The efficacy of case management services for severely mentally disabled clients. Community Mental Health Journal, 28, 163-180.

Stein, L. I., & Santos, A. B. (1998). Assertive community treatment of persons with severe mental illness. New York: W. W. Norton.

Stein, L. I., & Test, M. A. (1980). An alternative to mental health treatment. I: Conceptual model, treatment program, and clinical evaluation. Archives of General Psychiatry, 37, 392-397.

Stouthamer-Loeber, M., & van Kammen, W. B. (1995). Data collection and data management: A practical guide. (Vol. 39). Thousand Oaks: Sage.

Stroul, B. A. (1986). Models of community support services: Approaches to helping persons with long-term mental illness. Boston: Center for Psychiatric Rehabilitation.

Tanzman, B. (1993). An overview of surveys of mental health consumers' preferences for housing and support services. Hospital and Community Psychiatry, 44, 450-455.

Tarrier, N., Barrowclough, C., Vaughn, C., Bamrah, J., Porceddu, K., Watts, D., & Freeman, H. (1989). Community management of schizophrenia: A two-year follow-up of a behavioral intervention with families. British Journal of Psychiatry, 164, 501-506.

Teague, G. B., Bond, G. R., & Drake, R. E. (1998). Program fidelity in assertive community treatment: Development and use of a measure. American Journal of Orthopsychiatry, 68, 216-232.

Teague, G. B., Drake, R. E., & Ackerson, T. H. (1995). Evaluating use of continuous treatment teams for persons with mental illness and substance abuse. Psychiatric Services, 46, 689-695.

Test, M. A. (1979). Continuity of care in community treatment. New Directions for Mental Health Services, 2, 15-23.

Test, M. A. (1992). Training in Community Living. In R. P. Liberman (Ed.), Handbook of psychiatric rehabilitation (pp. 153-170). New York: Macmillan.

Test, M. A., Bond, G. R., McGrew, J. H., & Teague, G. B. (1997). P/ACT services research and fidelity to the model. Community Support Network News, 11(4), 5-7.

Test, M. A., & Stein, L. I. (1976). Practice guidelines for the community treatment of markedly impaired patients. Community Mental Health Journal, 12, 72-82.

Torrey, E. F., Erdman, K., Wolfe, S. M., & Flynn, L. M. (1990). Care of the seriously mentally ill: A rating of state programs. (3rd ed.). Arlington, VA: National Alliance for the Mentally Ill.

Torrey, W. C., & Wyzik, P. F. (1997). New Hampshire clinical practice guidelines for adults in community support programs. Lebanon, NH: West Central Services.

Trochim, W. M., Cook, J. A., & Setze, R. J. (1994). Using concept mapping to develop a conceptual framework of staff's views of a supported employment program for individuals with severe mental illness. Journal of Consulting and Clinical Psychology, 62, 766-775.

Unger, K. V. (1998). Handbook on supported education: Services for students with psychiatric disabilities. Baltimore, MD: Brookes.

Unger, K. V., Danley, K. S., Kohn, L., & Hutchinson, D. (1987). Rehabilitation through education: A university-based continuing education program for young adults with psychiatric disabilities on a university campus. Psychosocial Rehabilitation Journal, 10(3), 35-49.

Vogler, K. M. (1998). A fidelity study of the Indiana Supported Employment Model for individuals with severe mental illness. Unpublished dissertation, Indiana University-Purdue University Indianapolis, Indianapolis, IN.

Wallace, C. J., Liberman, R. P., MacKain, S. J., Blackwell, G., & Eckman, T. A. (1992). Effectiveness and replicability of modules for teaching social and instrumental skills to the severely mentally ill. American Journal of Psychiatry, 149, 654-658.

Waltz, J., Addis, M. E., Koerner, K., & Jacobson, N. S. (1993). Testing the integrity of a psychotherapy protocol: Assessment of adherence and competence. Journal of Consulting and Clinical Psychology, 61, 620-630.

Wang, Q., Macias, C., & Jackson, R. (1999). First step in the development of a clubhouse fidelity instrument: Content analysis of clubhouse certification reports. Psychiatric Rehabilitation Journal, 22(3), 294-301.

Ware, N. C., Tugenberg, T., Dickey, B., & McHorney, C. A. (1999). An ethnographic study of the meaning of continuity of care in mental health services. Psychiatric Services, 50, 395-400.

Webb, L. J. (1973). The therapeutic social club. American Journal of Occupational Therapy, 27, 81-83.

Winter, J. P., & Calsyn, R. J. (2000). The Dartmouth ACT Scale: A generalizability study. Evaluation Review, 24, 319-338.

Witheridge, T. F. (1991). The "active ingredients" of assertive outreach. New Directions in Mental Health Services, 52, 47-64.

Wolff, N., & Helminiak, T. W. (1996). Nonsampling measurement error in administrative data: Implications for economic evaluations. Health Economics, 5, 501-512.

Wood, R., & Steere, D. (1992). Evaluating quality in supported employment: The Standards of Excellence for Employment Support Services. Journal of Vocational Rehabilitation, 2, 35-45.

Woolf, S. H. (1990). Practice guidelines: A new reality in medicine. I. Recent developments. Archives of Internal Medicine, 150, 1811-1818.

Woolf, S. H. (1992). Practice guidelines: A new reality in medicine. II. Methods for developing guidelines. Archives of Internal Medicine, 152, 946-952.

Yalom, I. D. (1985). The theory and practice of group psychotherapy. (3rd ed.). New York: Basic Books.

Zahrt, D. M., Bond, G. R., Salyers, M. P., & Teague, G. B. (1999). Dartmouth ACT Fidelity Scale: A practical application in a statewide project (unpublished paper). Indianapolis: Indiana University Purdue University Indianapolis.

Zastowny, T., Lehman, A., Cole, R., & Kane, C. (1992). Family management of schizophrenia: A comparison of behavioral and supportive family treatment. Psychiatric Quarterly, 63(2), 159-186.