M. E. P. Seligman (1995) argued that traditional approaches to mental health services research fail to provide useful information to consumers and practitioners, particularly in an environment increasingly dominated by managed care. The authors recommend 4 guidelines for designing a research program so that the results can support the decisions of the major stakeholders (clients-families, practitioners, service managers, and policymakers): (a) Research must be targeted and programmatic and encompass a strategy of complementary efficacy, effectiveness, and cost-effectiveness studies; (b) study design and measure selection must be sensitive to who, and in what context, is best served by which intervention; (c) the design must inform stakeholders as to the type and amount of effort that is required to achieve a behavioral criteria; and (d) the strategy should inform researchers how information should be formatted to best support the decisions of the key stakeholders.

Seligman (1995) offered a provocative review of the limitations of our research methods in supplying stakeholders in the delivery of mental health services with information to support their decisions. Although many of the issues raised by Seligman are problematic, central in his argument are the burden and failure that the scientific community bears in supplying fruitful information to the consumers of our research, primarily the stakeholders of the mental health delivery system. We present and discuss four guidelines that should be considered in designing a program of research that generates information to support the decisions of the key stakeholders.

Stakeholders

Before presenting the four guidelines, we need to identify the key stakeholders and the types of decisions that a program of research could support. To do so, we will use the context of managed care, the most prevalent mechanism of service delivery of this decade. For the purposes of this discussion we consider four classes of stakeholders: (a) the consumer (including the client, the client’s family, and the party paying for the service), (b) the practitioner, (c) the supervisors and service manager(s), and (d) the policymaker who sets the standards for care and reimbursement. The primary set of stakeholders are direct consumers of the service (the client or patient) and the consumer’s family or employer (who are directly or indirectly paying for the costs of the services). Consumers ask themselves, “Will I be able to function adequately. . . . What is required of me and can I afford the services?” The practitioner(s) providing the services, the clinical supervisors, and managers of the service want to understand, “What interventions will have optimal impact on consumers of a specific psychosocial profile in the bounds of the resources that have been made available to us?” Those persons who set policies for quality and reimbursement standards want to understand, “What outcomes, processes, and cost-constraint standards are feasible for each class of consumers?” An additional stakeholder group that would influence publicly funded services are those agencies and agents across the community who are indirectly impacted by delivery of such services (e.g., health, legal, or social service providers). They want to understand, “What is our role or responsibility to or for persons with a mental health problem?”

Currently, most insurance carriers and managed care organizations are following two lines of evidence to direct their policies: cost containment and consumer satisfaction. Few organizations have used behavioral criteria beyond gross diagnostic categories to determine service eligibility or the level of service that is reimbursed. The amount and type of service are often based on a rule of dosage (e.g., 12, 18, or 20 sessions of psychotherapy; 10, 21, or 30 days of inpatient service; or both). Consider applying the logic of these cost containment rules in oncological medicine:

“Mr. Smith, 90% of the tumor is now benign. Only 10% of the tumor remains malignant. Unfortunately you have used up the 20 sessions of radiation treatment allowed under your managed care plan for this year. Please come back next year when your insurance eligibility has been renewed.” (Newman & Carpenter, in press)

Frederick L. Newman, Health Services Administration, Florida International University and Center for Family Studies, Psychiatry and Behavioral Sciences Department, University of Miami; Manuel J. Tejeda, Center for Family Studies, Department of Psychiatry and Behavioral Sciences, University of Miami.

Five colleagues contributed greatly to this article: Jeff Berman, Gary Bond, Will Shadish, Gregory Teague, and Brian Yates.

Correspondence concerning this article should be addressed to Frederick L. Newman, HSA-SPM, Florida International University, North, 3000 NE 145th Street, North Miami, FL 33181. Electronic mail may be sent via Internet to newmanf@servms.fiu.edu.
The low correlations between indexes of consumer satisfaction and actual impact on functioning or quality of life clearly indicate that this approach, albeit currently popular, is limited in its usefulness for guiding consumers in the selection of a service provider.

Whether they are privately or publicly operated, managed care organizations are seen as holding only profit or cost constraints as their guiding principle. However, there is strong evidence that these organizations cannot survive unless they consider both the effectiveness (outcomes) and the costs of providing the therapeutic interventions available under their managed care plan. For long-term survival in a complex industry such as health care, managed care plans must consider the outcomes and the costs to three groups: (a) the payer (including the employer, or the government, and the copaying consumer or family), (b) the service provider, and (c) those in the community affected by the client-patient's behavior.

Our research strategies are often delinquent in addressing these issues. The consumers of our findings inform us of this failure. Thus, we argue that our research designs must begin to incorporate the needs of our consumers, specifically to support the decision making of all parties involved in the delivery of mental health care.

An additional group of stakeholders is the scientific community itself, who advance and test basic theories about behavior and the psyche. This group cannot be neglected here—we are that group! Our colleagues are our consumers, and we must urge ourselves to operate under the same ideological umbrella voiced by Bronowski (1959), who recommended that science must be guided by the simultaneous exploration and examination of human values. An underlying value is that our science should provide information useful to each of the four sets of decision makers. In this context, we offer the following set of guidelines for designing a program of research that is intended to support the decisions of these stakeholders.

**Efficacy, Effectiveness, and Cost-Effectiveness**

Research must be targeted, programmatic, and encompass a strategy of complementary efficacy, effectiveness, and cost-effectiveness studies. This first guideline recommends that a broad but targeted program of research be envisioned to develop and test the efficacy and effectiveness of a therapeutic intervention. Randomized controlled trials are typically needed in the early stages of development, not only to refine a technique that may produce stronger effects but also to permit testing for the intervention's safety under controlled conditions.

It is interesting that one requirement of a well-designed efficacy study is to institute monitoring methods to assure treatment fidelity—who does what with which, when, where, and to whom are obsessively documented. These data are precisely the basic ingredients required for many forms of cost-effectiveness research (see Knapp, 1995; Yates, 1996). Once a reasonably strong and safe intervention has been demonstrated by the efficacy studies, a program of effectiveness and cost-effectiveness studies can help us understand who is best served by the therapeutic intervention and in what environmental context. To be sure, both efficacy and effectiveness studies can serve the same master provided there is a clear understanding of the overall mission of a research program and how complementary strategies can benefit that mission.

Although the meta-analysis of meta-analyses by Lipsey and Wilson (1993) discovered that randomized control trials and quasi-experimental (effectiveness) studies yield similar effect sizes, Shadish and his colleagues (Heinsman & Shadish, 1996; Shadish & Ragsdale, in press) found that randomized trial studies yielded stronger effect sizes. These differences were explained by the fact that effectiveness studies typically have less control over facets that increase error variance (the denominator of an effect-size estimator).

Seligman (1995), among others, correctly noted that randomized control trials and well-controlled effectiveness studies have advantages that the practicing clinician does not. Specifically, the practicing clinician does not have full control over who walks in the door and has less control over spurious events between sessions. Nevertheless, important issues at both ends of the control spectrum can be addressed from the continuum of research designs—from efficacy studies, through the nine-plus quasi-experimental designs, to the retrospective surveys represented by the Consumer Reports (CR; 1994) survey. Let us compare the differences in the interpretations that come from the CR survey with that of the Shadish and Ragsdale (in press) meta-analysis as a case study that argues for the need for complementary efficacy and effectiveness research. In the CR survey, marital and family counselors were not found to be as effective as psychiatrists, psychologists, or social workers. Yet the Shadish and Ragsdale (in press) meta-analysis on 64 efficacy studies (with random assignment) and 36 effectiveness studies (using non-equivalent group designs) found what are often characterized as large effect sizes for family psychotherapy. Note that marital therapy was not included in their meta-analysis. They also found that the effect sizes were significantly larger for the efficacy studies (with random assignment) than for the quasi-experimental (nonequivalent group) studies.

More research is needed to partition the effects of the type of therapy from the therapist's discipline to better understand the discrepancy between the clear success of the 100 marital and family therapy studies in the Shadish and Ragsdale (in press) meta-analysis and the findings in the CR (1994) survey. The CR survey results appear to provide the first major evidence that training and licensure requirements do make a difference, whereas other single studies or meta-analyses have not shown the same results (Berman & Norton, 1985). Moreover, Berman (personal communication, February 16, 1996) recommended that "any observed differences between disciplines in the CR survey are likely to be differences in the type, problem severity, and prognosis of clients who choose different types of helpers."
Should we interpret these results to mean that when marital or family therapy is used in the real world, it is not effective? Or are there alternative explanations, as suggested by Berman (personal communication, February 16, 1996), Shadish and Ragsdale’s (in press) results suggest that preselection and facets influencing both internal and external validity strongly affect the predicted outcome. Furthermore, it should be obvious that marital and family therapy is one area that calls for a series of effectiveness studies to help us understand what real-life circumstances could negatively influence the potential positive effects of a therapy with demonstrated efficacy.

Although the CR (1995) article raised several issues that will guide follow-up research, its bold conclusions about family and marital therapy appear to be too broad, and the reporting of the data could have cautioned the reader as to the potential for this misinterpretation. There are notable examples in the scientific literature in which a series of well-designed randomized controlled trials questioned basic assumptions and were key to identifying more appropriate intervention strategies. If popular beliefs were allowed to prevail, these bolder intervention strategies might not have had a chance to show their potential. To provide a good example of this, we asked Bond (personal communication, February 19, 1996) to identify random assignment studies that were, and still are, seminal to understanding how to effectively deliver mental health services to persons with a severe and persistent mental illness in the community. His recommendations provided a strong argument for randomized control trials in an area in which there has been little belief that community treatment is either efficacious or cost-effective. We quote directly from Bond’s electronic mail to us:

May, Tuma, and Dixon (1981), examining inpatients with schizophrenia, compared five treatment conditions: psychotherapy only, psychotherapy and phenothiazine, electroconvulsive shock therapy, milieu, and psychotherapy plus medication. The psychotherapy-only condition had the worst outcomes overall, similar to outcomes for those receiving milieu treatment only. This study pretty much dispelled the idea that psychotherapy as practiced was the best treatment for schizophrenia. Katon and VandenBos (1981) questioned whether the “right” form of psychotherapy was provided and did find that psychotherapy tailored to the population was shown as cost-effective.

Hogarty, Goldberg, Schoeler, and Ulrich (1974) shifted the focus of research to that of combined medication and an array of psychosocial interventions rather than just psychotherapy. Hogarty et al. performed a randomized controlled trial of psychotropic medication and major role therapy for people with a diagnosis of schizophrenia. They found that psychotropic medication had a major effect in reducing relapse, but that psychosocial interventions also had an incremental effect above and beyond what the medication contributed. From the late 1970s through the early 1990s, a series of randomized clinical trials employing assertive community treatment which integrated medication management with different forms of psychosocial interventions were performed by Stein and Test (1980) and by Bond and his colleagues (see reviews in Bond, 1991, 1992; Bond, McGrew, & Fekete, 1995). These teams of researchers repeatedly demonstrated the efficacy of these models to increase community tenure and decrease number of days of hospitalization.

Until Weisbrod’s (1983) economic analysis of the Stein and Test (1980) study, most people believed that the model of assertive community treatment programs had hidden costs that actually made them more expensive than episodic treatment approaches. Weisbrod had a fairly extensive accounting framework applied to a randomized controlled trial. He showed that a model program had about the same cost as controls, and with better client outcomes. A clinical trial by Drake, McHugo, Becker, Anthony, and Clark (1996) dispelled the myth that people with severe mental disorder need a period of care exploration before they are ready to start employment. (Bond, personal communication, February 19, 1996)

Taken together, Bond (personal communication, February 19, 1996) argued that a series of randomized control trials over the last three decades has focused our attention on the possibility of cost-effectively serving our most persistently impaired citizens in the community with a possibility of community employment. But one may still ask whether this is true of all subsets of people with a history of persistent and severe mental illness. Are there permutations on the basic model that might serve some subgroups better? Here the issue concerns identifying necessary ingredients in assertive community treatment that may have their most potent impact on different types of consumers. What ingredients are necessary for all groups?

McGrew, Bond, Dietzen, and Salyers (1994) performed a meta-analysis across 18 sites on assertive community treatment and identified those ingredients that were associated with the greatest number of days reduction in hospitalizations. Does that retrospective analysis set the stage for a new set of prospective studies in which the ingredients are systematically varied, or should we simply use the results of the retrospective analysis to identify the most effect ingredients? Would those supporting the view of outcome represented in the CR (1995) article recommend that we simply accept the retrospective analysis performed by McGrew et al.? Probably not, once someone points out that some of the ingredients (e.g., a 1:10 staff-to-consumer ratio and that both a medical doctor and a nurse should be on the team) are expensive.

Finally, treatment- adherence data collected in most efficacy studies contain the basis for cost-effectiveness analysis. In its most basic form, treatment adherence is nothing more than the managerial accounting of the behaviors of individuals involved in the delivery of service. Moreover, accounting for the behaviors of individuals is not only sound science but effective management practice ( Mintzberg, 1973). It requires little more effort in the management of a scientific endeavor to account for the activities of clinical staff on a scientific project. Assessing staff time and whereabouts provides the fundamental pieces for a cost-effectiveness analysis. As argued over a decade ago (Newman & Howard, 1986), our traditional procedures of collecting data on assuring treatment integrity and assessing therapeutic effort in our research studies are sufficient to perform some of the most basic cost-effectiveness studies. Moreover, the procedures described by Knapp (1995) and his colleagues by either a
structured interview or a self-report are sufficient to collect the information needed for more sophisticated cost-effectiveness studies. Stated quite simply, the technology exists for us to collect the data to conduct cost-effectiveness analysis even with our most basic efficacy studies and certainly with our broader naturalistic studies.

The story here is one in which a combination of randomized control trials, effectiveness, and retrospective studies have come together under a program of research. Given the evolution of this research, it is probably time to perform a CR (1994)-type survey with the consumers of assertive community treatment services. As described later, several such efforts are now in progress. But there is an intermediate issue to be discussed first.

Who Is Best Served and in What Context?

The study design and measures selected must be sensitive to produce results that inform us about who, and in what context, is best served by which intervention. This second guideline recommends that this can be done in a manner that is ultimately useful to the four classes of stakeholders only if the program of research systematically considers what attribute and contextual moderator variables will influence the effectiveness of the intervention. There are a number of moderator variables that should be helpful to both the consuming–paying public and the professional communities in the choice of therapeutic mode or clinician. At a minimum, the study design should include the analysis and report of the interaction (covariation) of the initial level of symptom distress, functional status, or self-management ability with treatment effects.

Although the CR (1995) article did attempt to describe some of the interactions, the questionnaire (CR, 1994) design and the analyses used did not permit a thorough exposition of who is best served by which intervention to be addressed. Those involved in the design and analysis of the survey are not to be faulted. Creating a study design and selecting measures that are sufficiently reliable and sensitive to detect such interactions are profoundly difficult, with a great deal of debate in the literature as to what can and should be done. What follows is an analysis of the problems with some recommendations.

An important issue embedded here is whether one should separate out those variables that would be important to consider when designing how a therapeutic intervention is delivered so that it has its greatest impact. An approach often used by some researchers is to statistically adjust for differences observed among participants after treatment for those differences that were found prior to administering the treatment. Covariance adjustments or residual gain scores are two of these approaches (Cronbach, 1992; Cronbach & Furby, 1970). The intent of this adjustment is to estimate treatment effects independent of differences in initial status. A variation on this argument is that one can more cleanly test a theory that predicts between group differences if such pesky initial variations are taken out of the equation (the error term). Some of the analyses in the CR (1995) article used such adjustments, and some methodologists would question this approach.

An argument against using adjustments for initial differences in a test of between-group differences is one of basic relevance (Newman, 1994; Willett & Sayer, 1994). If there exists a common source of covariation in the day-to-day delivery of services, then that source of the covariance should be studied rather than subtracted out of the analysis. Specifically, a program of research on the effectiveness of a particular intervention should include the major sources of variation that one might reasonably expect in the delivery of the services in the field. If between-group differences are found to be significantly above an acceptable noise level represented in current practice, then these differences are very much worth knowing.

There are several problems with this view when it is taken to the extreme. Often, an innovative approach requires careful development in its early stages. Early development and test work probably should be discovery based (Greenberg & Newman, 1996), and early discovery-oriented exploration should be followed by research designs that are most sensitive to the intervention’s potential benefits. Any known technique that can control error variance should be part of the designs at these early developmental stages. However, it is critical that the program of research lead to testing the robustness of the intervention under those conditions that are common in the field of application.

The specific recommendation is that once the research program achieves a level in which effectiveness studies are indicated, then the sources of covariation that are due to key moderator variables, including that of the initial status, must be studied and not statistically controlled (removed) in the analysis and reports. Including these interactions in the study design will move us toward the goal of identifying for whom a therapeutic intervention is best suited and away from the do-do bird interpretation that all the therapies are winners.

There are two traditional problems that arise when studying interactions between treatment and moderator variables. One is inherent in our methods of measurement (i.e., reliability), and the other is whether our methods of analysis have sufficient power to detect effects. Shoham-Salomon (1991; Shoham-Salomon & Hannah, 1991) and Beutler (1991) have documented the poor record of isolating such interactions in psychotherapy research. In those instances in which interactions have been found, the researchers had a strong theoretical reason for selecting the moderator variable as well as the dependent measures. In those cases in which generic measures of outcome were used, significant results typically were not obtained. One inference from their reviews is that the outcome measures need to be more carefully selected when studying an interaction effect. A more important inference is that theory should not be underestimated.

Simulation studies by McClelland and Judd (1993) and by Jaccard and Wan (1995) have provided ample
demonstration that detecting interactions and moderator effects typically have low power even when difficulties that are due to measurement reliability are minimal. When measurement reliability is a problem, the issue of power worsens. Aiken and West (1993) provided an excellent primer on how to deal with testing and interpreting interactions through regression techniques.

Recent findings have been more encouraging. Jaccard and Wan (1995) found that structural equation approaches are more powerful in detecting such interactions. Even more exciting are the findings by Willett and Sayer (1994) that structural equation modeling can also be used to test differences in patterns of change among different subgroups within and between treatment groups. Willett and Sayer ended their article with the strong recommendation that "we can measure 'change'—and we should" (p. 379).

Several authors (Bryk & Raudenbush, 1988; Lyons & Howard, 1991; Willett & Sayer, 1994) described procedures that can identify when a source of error variance may be masking the interaction of a moderator and a treatment variable that should be investigated in follow-up research. However, as indicated earlier, one alternative explanation for large error variance remains measurement reliability–unreliability (Lyons & Howard, 1991). Another is that neither the main effect of treatment nor the interaction effects exist or are sufficiently large to matter.

Another problem in the description of key interaction effects is our field's questionable allegiance to statistical significance (Cohen, 1994; Schmidt, 1996b). The politics of p values are ubiquitous in the industry of science. Our typically underpowered studies are likely to not be published in refereed journals or any journal, for that matter. Although this issue is recognized by editors, we have not found a satisfactory substitute to guide the accepting or rejecting of research reports in our scientific archival journals.

One possible sign is that Division 5 (Evaluation, Measurement, and Statistics) of the American Psychological Association has formed a work group to study the issue with other scientific associations, with an objective to formulate guidelines for editorial review that are less reliant on statistical significance (Schmidt, 1996a). Another solution may lie in using effect sizes. Although effect sizes, power, and p values are interlinked, small samples can produce low power and nonsignificance while yielding a modest effect size. Overdependence on statistical significance in small samples while ignoring effect sizes may lead us to ignore relationships that may be important—the Type II error. Science has yet to deal effectively with this issue: Unlike Type I errors in which the probability can be computed exactly, the precise probability of Type II errors is usually, and uncomfortably, unknown.

The bottom line here is that a program of research should plan to describe the interaction effects of a treatment with the principal moderator variables that clinicians face in their daily practice. Although the literature has noted that there are measurement, design, and analysis techniques that have the potential to do this, an unmet challenge is whether retrospective surveys, such as the CR (1994) survey, can be designed to be sensitive to these interaction effects. We need to improve on our ability to design prospective studies that are sufficiently sensitive to study such interactions, so that a parallel technology will emerge for retrospective surveys.

Behavioral Criteria Must Drive Research Design

The study design must inform us as to the type and amount of effort required to achieve behavioral criteria rather than continuing to perform studies that fix the dosage of the intervention. This third guideline recommends that the design of our studies must focus more on what efforts and resources are necessary and sufficient to attain specific levels of outcome. Specifically, we must seek to redesign our research questions, hypotheses, and methods so that we can estimate the type and amount of effort needed to attain a given range of behavioral outcomes. This requires that we move away from studies that fix dosage (i.e., number of sessions) and move toward designs that seek to describe the type and amount of effort likely to attain specific levels (ranges) of behavioral outcome.

There is a need to investigate the relationships between efforts (costs) for an intervention to attain a desired behavioral outcome criteria rather than continuing to perform studies that fix the dosage of the intervention. One way of summarizing the type and amount of efforts required by an intervention strategy is by estimating its costs. Costs should be seen as an estimate of the efforts or resources consumed by the one or more parties involved directly or indirectly with the therapeutic intervention. What resources are expended by the client, the service provider—therapist, the family-support system, or the community—society? There is a homily that describes the cost-effectiveness issue for the service provider—therapist: Every decision made about a client's clinical or functional status is related to resources (and their costs) needed to help the person. Likewise, every decision made about the allocation of resources (and their costs) is related to the potential impact a service can have on a client's clinical status—functioning. We hope that our science will provide adequate cost-effectiveness data to support the decisions about allocating the resources needed to achieve desired outcomes.

The technology to estimate costs of resources and relate these to outcomes has been in our literature for quite some time (Carter & Newman, 1975/1980; Newman & Howard, 1986; Yates & Newman, 1980a, 1980b) and has been updated (Clark et al., 1994; Knapp, 1995; Yates, 1996). However, few research groups have considered resources used (through costs or dosage) as a systematic feature of their research program. Four groups do stand out as potential role models. The research groups at the University of Wisconsin—Madison and the Indiana University—Purdue University at Indianapolis have focused on assertive community treatment (Bond, 1991; Weisbrod, 1983) for persons with a persistent and severe men-
tal illness. The research group at Dartmouth-New Hampshire Psychiatric Research Center (Clark et al., 1994; Drake et al., 1996) has focused on persons with a dual diagnosis of a severe and persistent mental illness and substance abuse. Howard’s (Howard, Moras, Brill, Martinovich, & Lutz, 1996) group at Northwestern University has focused on persons seeking psychotherapy, and these have primarily been persons with an affective disorder or a situational crisis.

Another technology under development is that of linking level of care to level of need (Carter & Newman, 1975/1980; Leff, Graves, Natkins, & Bryan, 1985; Leff, Mulkern, Lieberman, & Raab, 1994; Newman, Griffin, Black, & Page, 1989; Uehara, Smukler, & Newman, 1994). This technology, as currently understood, assumes that a necessary first step is to perform a holistic assessment profile of a consumer population that partitions the population into subgroups that have homogeneous needs for service resources (cost-homogeneous groupings). The profile is described as holistic because it considers physical needs along with psychological, social–interpersonal, and community-functioning needs. Those who work with programs that serve persons with a severe and persistent mental illness have had to face the need for this holistic view in designing and evaluating mental health services since the early 1970s when deinstitutionalization began (Carter & Newman, 1975/1980; Leff et al., 1985). A more specific description of a cost-homogeneous group is one that describes persons with a similar physical and psychosocial profile as requiring similar service needs to achieve a given set of behavioral criteria in a given timeframe. It should be noted that, for some cost-homogeneous groups, the behavioral criteria may be to maintain functioning in a given range. For others, the criteria might be to improve functioning.

An idealized example describing how behavioral criteria can be used to evaluate the relative cost-effectiveness of two or more groups would be useful at this point. This example uses technology that is currently available (with one limitation that we discuss later). The intended message in reviewing this example is that by using existing technology, one can describe the impact of two or more interventions in terms of its cost-effectiveness in achieving observable criteria. The example illustrated in Figure 1 offers a graphical analysis in the form of a progress–outcome report for three hypothetical groups of individuals. Persons in these three groups have been identified as having similar characteristics in their cost-homogeneous grouping. In this example, all three groups have similar initial levels of functioning and have service plans for a 26-week period (182 days), so that the expected costs of administering the treatment plan are the same. However, two groups differ with regard to their behavioral goals. Group 1 is expected to improve in their overall functioning to a point in which they can be placed in another cost-homogeneous grouping requiring a less intensive–expensive service plan. Group 2 has a goal of maintaining their functioning in the bounds described as acceptable to community functioning. Group 3 represents outliers from Group 2 and is discussed later.

The vertical axis of Figure 1 represents a person’s overall ability to function in the community. It is understood that this global measure must be supported by a multidimensional view of the persons in each group (Newman & Ciarello, 1994). The horizontal axis represents the cumulative costs of providing services from the beginning of this episode of treatment. The two horizontal dotted lines inside the box represent the behavioral criteria set a priori (i.e., the lower and upper bounds of functioning for which this mental health service is designed to serve). If a consumer is observed to behave at a level below the lower dotted line, then she or he should be referred to a more intensive service. Likewise, if a consumer is observed to behave at a level above the upper dotted line, then either services should be discontinued or a referral to a less intensive–expensive service should be considered. Finally, the dashed vertical line inside the box represents the cumulative costs of the planned services for the 26-week period. In this hypothetical case, all three groups have treatment plans with the same expected costs.

The circled numbers in Figure 1 track the average progress of consumers in each of the groups over 13 successive 2-week intervals (a total of six months). The vertical placement of the circled number represents the group’s average level of functioning at that time, and the horizontal placement represents the group’s average costs of services up to that point. The sequence of the 13 circled numbers represents the average progress of persons in each of the groups in 2-week intervals over the 26 weeks of care. For Group 1, with a six-month objective of improvement above that of the upper dotted line, the objective is met. For Group 2, with the six-month objective of maintaining functioning within the range represented by the area between two dotted lines, the objective is also met. Group 3 is an example of clients who exceeded the bounds of the intended service–treatment plan after four months, two months shy of the objective. These consumers were able to maintain their community functioning but only with the commitment of additional resources. A continuous quality improvement program should focus a posthoc analysis on these consumers to determine if they belong to another cost-homogeneous subgroup or whether the services were provided as intended and whether modification of current practices are needed.

A major difficulty in attempting to enact these recommendations is that of obtaining a believable database to determine what the appropriate progress–outcome criteria or what the expected social or treatment costs should be. Some have used expert panels to set the first draft of such criteria (e.g., Newman et al., 1989; Uehara et al., 1994). Howard et al. (1996) have used baseline data from prior studies, along with measures taken on a nonpatient population. However, both strategies have problems. The research literature is largely based on studies in which the dosage was fixed within a study, thereby constricting inferences about what type and amount of
Figure 1
Hypothetical Example of a Graphical Analysis of Changes in Functioning and Costs Relative to Managed Care Behavioral Criteria for Three Cost-Homogeneous Groups of Consumers

Note. Group 1 had the planned objective of improvement in functioning so that they could move to a less intensive service plan within six months. Group 2 was expected to maintain functioning over the same six months but to use about the same amount of resources. Group 3 was an outlier group who required additional resources to maintain adequate levels of community functioning. They are highlighted here for a follow-up study by the program’s quality improvement team.

effort can achieve specific behavioral criteria. Howard’s early work on a dosage and phase model of psychotherapy was statistically confounded by combining the results of controlled studies in which number of sessions was fixed. Some of the naturalistic data came from persons who had no limit on sessions, and others did have limits on the number of sessions, which was imposed by third-party payers. Having worked closely with Howard over the last decade, it would appear the dosage data as reported are probably trustworthy. However, we know of no studies in which behavioral criteria were set a priori, and the type and amount of efforts were seen as the dependent variables in estimating the success or failure to achieve these criteria. Yet the logic of managed care and the logic of practice guidelines under discussion by the American Psychological Association and the American Psychiatric Association would require that such study results be available to set criteria for using a particular intervention strategy or to set reimbursement standards. This void must be filled by data from well-designed efficacy and cost-effectiveness studies that can provide empirical support for setting behavioral outcome criteria for managed care programs. Without such data, there is faint hope of changing the current practice of setting dosage guidelines independent of behavioral criteria.

How would one go about establishing such criteria? The design of a project currently underway for the Indiana Division of Mental Health represents our understanding of what the literature recommends in the constraints of the real world of Indiana. The project is designed to provide an empirical base for a team of actuaries to set these criteria approximately 2.5 years

1 Richard DeLiberty, deputy director of mental health, is the overall project director and directs the consumer telephone survey project. Kay Hodges of Eastern Michigan University directs the development of the assessment instrument for children and adolescents. John McGrew of Indiana University—Purdue University of Indianapolis directs the field research team activities. Manuel J. Tejeda provides statistical consultation. Frederick L. Newman leads the development of the adult assessment instrument and directs the technical aspects of the project.
Research on Communicating With All Stakeholders

How must information be formatted to best support the decisions of the key stakeholders (clients–families, clinicians, service managers, and policymakers)? This fourth guideline serves to remind the scientific community that research cannot exist in a vacuum: We have a tendency to talk to ourselves. Scholars must learn to communicate their results effectively to all consumers of their research and not simply to one another. We must keep in view each of the key stakeholder groups when exploring and designing how we will communicate our findings. Here, we are indebted to scholars such as Seligman (1995) and Howard and his colleagues (1996) for furthering diagnosis and treatment as scientists, while exploring new ways to test and communicate the findings of the research to those concerned with the delivery of services.

There is a need to perform studies on the use of effectiveness and evaluation data in supporting managed care decisions by the stakeholders. As described earlier, there are at least four sets of stakeholders when considering the managed care environment in its broadest sense: (a) the client or family in their choice of modality or clinician, (b) the clinical staff in supporting their selection of therapeutic strategy, (c) the management staff in the allocation of resources that meet the needs and can potentially achieve the behavioral outcome criteria, and (d) those who set funding or accreditation policy about the linkage between level of need and level of resource availability in a managed care environment. Whether it is the family, the clinician, the employer, the insurance carrier, or a government entity, the central issue is how to best manage the available resources to meet the need and achieve a desired outcome. Study designs must assess resources used with the traditional measures of initial status, and progress and outcome should support client–family, clinical, supervision, management, and policy decisions in a managed care environment.

There has been a start in the development of consumer-oriented report cards on the access, availability, outcome, and satisfaction with mental health services. Mulkern, Leff, Green, and Newman (1995) reviewed the published and unpublished materials that were available in the spring and winter of 1995. The technology of developing performance indicators on accessibility, availability, and satisfaction does exist. Kamis-Gould (1978) and Sorensen and his colleagues (Sorensen, Zelman, Hanbery, & Kueic, 1978) developed a battery of performance indicators to describe access, availability, and financial stability viability of mental health services. Davies and Ware (1991) developed a measure for the consumer satisfaction for the Group Health Association of America that has reasonable psychometric properties.

Outcome measures from the viewpoint of the consumer are not readily available. Yes, there are many outcome measures that are described as consumer self-reports (e.g., the SF-36 [Ware, Snow, Kosinski, & Gandek, 1993] or the SCL-90 [Derogatis, 1975]), but these have
been derived by mental health professionals to estimate impact from the viewpoint of a mental health professional. Measures derived from consumers' views are still being developed and tested for their psychometric quality (Mulkern et al., 1995).

The Center for Mental Health Services (CMHS) created a task force of consumer representatives and program evaluators to develop such an instrument under the Mental Health Systems Improvement Program (MHSIP). The task force reviewed all of the instruments available and gathered information from a variety of consumer focus groups (some using fairly sophisticated multidimensional scaling procedures). The task force drafted an instrument that is now available. A number of states, including Indiana, are attempting to develop and test the use of their own mental health services report card, using the CMHS model as a guide.

The concept of packaging research and evaluation findings for the principal sets of stakeholders is not currently a focus of our science or seen as the concern of the clinical services researcher. Yet it is central to the success of our field having impact on managed care guidelines. It may be time for us to review the research performed by those in decision sciences, management, and marketing to better understand how information can be packaged so that the information is useful to consumers and their families, practicing clinicians and their administrators, and policymakers influencing the rules of managed care.

Conclusion

The CR (1995) article and Seligman (1995) are to be credited for their provocations. Our intentions have been to share some of the limitations we and others have observed in the state of the science. Our work as scholars has made great inroads in treatment. However, the changing and turbulent environment of health care now demands a broader vision.

We must contrast the broader vision with the demand for specificity that is an integral part of the social sciences. Indeed, our studies and careers reflect the narrow vision required in our research. The standard recommendation in our texts (e.g., Meyers & Well, 1991) is that we should seek to maximize systematic variance, minimize error variance, and control extraneous variables. As we strive to create designs of greater internal validity, the consumers of our research become increasingly less central.

Our studies, although scientifically thorough, typically are not devised to assess the real-world impact of our interventions. Our science does justify certain forms of treatment, but we traditionally fail to appreciate the managerial, legal, and social ramifications resulting from endorsing one intervention over another. Although we cannot be expected to assess every repercussion of our research, we can begin to include in our research the components that begin to help us assess some of these repercussions.

Too often, mental health care decisions are made because of fiscal constraints. In part, we are to blame. On the one hand, the costs of mental health services are testament to the increasing complexities of providing services. On the other hand, the technologies have existed for some time for us to incorporate therapeutic effort, level of care, and cost information in our research (Newman & Howard, 1986). We have arrived in times in which technologies not only exist but also in which applications are required to simultaneously answer the questions of science, clients, providers, managers, and policymakers. This article has followed that theme: Technology now exists at such a level that we can adapt our techniques to the information demands of the stakeholders in the (managed care) environment of mental health service delivery. If we return to one question we proposed in the beginning of the article, the client-consumer is still asking, "Will I be able to function adequately, . . . What is required of me, and can I afford the services?" Our studies most move to answer, "Yes, and this is how."

REFERENCES


October 1996 • American Psychologist