Fidelity Criteria: Development, Measurement, and Validation

CAROL T. MOWBRAY, MARK C. HOLTER, GREGORY B. TEAGUE, AND DEBORAH BYBEE

ABSTRACT

Fidelity may be defined as the extent to which delivery of an intervention adheres to the protocol or program model originally developed. Fidelity measurement has increasing significance for evaluation, treatment effectiveness research, and service administration. Yet few published studies using fidelity criteria provide details on the construction of a valid fidelity index. The purpose of this review article is to outline steps in the development, measurement, and validation of fidelity criteria, providing examples from health and education literatures. We further identify important issues in conducting each step. Finally, we raise questions about the dynamic nature of fidelity criteria, appropriate validation and statistical analysis methods, the inclusion of structure and process criteria in fidelity assessment, and the role of program theory in deciding on the balance between adaptation versus exact replication of model programs. Further attention to the use and refinement of fidelity criteria is important to evaluation practice.

INTRODUCTION

Effectiveness research is now at the point of sophistication wherein black-box outcome studies are no longer acceptable (Chen, 1990; Yates, 1994, 1995, 1996). Rather, interventions are expected to specify the model—a scientifically sound program theory or theory of action, explicating the mechanisms through which the program will achieve its desired outcomes. There should be valid and reliable criteria for establishing fidelity to the model, as well as specific treatment inclusion criteria (Bond, Evans, Salyers, Williams, & Kim, 2000; Hohmann & Shear, 2002; National MH Advisory Council, 1999). High quality treatment effectiveness research should utilize a program manual for training and supervising intervention staff and for monitoring program quality and performance, helping to ensure fidelity to the model being researched.

Carol T. Mowbrey • School of Social Work, University of Michigan, 1080 South University, Rm 2734, Ann Arbor, MI 48109-1106, USA; Tel: (1) 734-763-6578; Fax: (1) 734-763-3372; E-mail: cmowbray@umich.edu.

American Journal of Evaluation, Vol. 24, No. 3, 2003, pp. 315–340.
ISSN: 1098-2140

Thus, the development and use of valid fidelity criteria is now an expected component of quality evaluation practice.

In addition, the significance of fidelity criteria is now even greater, given the current emphasis on the need to utilize evidence-based practices, which is increasingly being brought to the attention of professionals and the public (IOM, 2001). Consumers and family members see access to treatment practices that are strongly supported by research as an appropriate expectation of any system of care in the human services (Drake et al., 2001; Frese, Stanley, Kress, & Vogel-Scibilia, 2001; NH-Dartmouth Psychiatric Research Center, 2002). Researchers and evaluators should, thus, be concerned about the extent to which their work contributes to determinations as to whether an intervention, service, or program model that they study is evidence-based or not. If the research or evaluation does not come up to the highest possible standards of evidence, then its usefulness is obviously compromised. However, beyond encouraging randomized clinical trials (RCTs), the literature on research and evaluation methodologies does not appear to well specify exactly what the expectations are for research quality and design to maximally contribute to establishing an evidence-base for a given program. Clearly such expectations go beyond the use of randomized clinical trials. Also, there are many situations in which RCTs are either not possible or not practical (e.g., systems-level designs which involve interventions at the level of cities, counties, or other large governmental units). The appropriate use of fidelity criteria can assist program evaluation designs, with or without RCTs, to contribute to establishing the evidence-base for any program.

This article is intended to improve understanding and articulate important evaluation issues related to fidelity criteria and their relevance to the evidence-based practice literature. The audience is intervention and services researchers and program evaluators in health, education, and human services. In this article we review and discuss literature on and applications of fidelity criteria: What they are, why they are important, and how to develop, measure and validate them.

What is Fidelity?

Fidelity is a relatively recent concept in some arenas (e.g., mental health services research), although its use in program evaluation can be dated back 20–25 years (Sechrest, West, Phillips, Redner, & Yeaton, 1979). Blakely et al. (1987) cite pioneering work by Hall (unpublished) which described social programs as consisting of a finite number of components and defined fidelity as the proportion of program components that were implemented. Bond, Evans, et al. (2000) provide a brief history of the development and use of fidelity measures in the mental health field, starting with research on psychotherapy outcomes. In outcome research, fidelity has been described as "confirmation that the manipulation of the independent variable occurred as planned" (Moncher & Prinz, 1991, p. 247). Fidelity assessments have administrative as well as research purposes, for example, to determine how adequately a program model has been implemented (Bond, Evans, et al., 2000), to assess conformity with prescribed elements and the absence of nonprescribed elements (McGrew, Bond, Dietzen, & Salyers, 1994), or to provide assurances to policy-makers that services are being implemented as intended and are reaching the target audience (Orwin, 2000). In non-research contexts, fidelity may be simply defined as "the adherence of actual treatment delivery to the protocol originally developed" (Orwin, 2000, p. S310). Typically, scales are developed to quantify fidelity—to compare a programmatic clinical intervention, as implemented, to the empirically tested model on which it is based (Drake et al., 2001).

Why Assess Fidelity?

Fidelity criteria, used as a manipulation check in treatment effectiveness research, are necessary to ensure internal validity (Hohmann & Shear, 2002). Practically speaking, the most oft-cited reason for assessing fidelity is the need to account for negative or ambiguous findings (Hohmann & Shear, 2002). That is, without documentation and/or measurement of a program's adherence to an intended model, there is no way to determine whether unsuccessful outcomes reflect a failure of the model or failure to implement the model as intended (Chen, 1990). In fact, failed implementation is the most common reason for failed outcomes according to some sources (Mills & Ragan, 2000). Bond, Evans, et al. (2000) indicate that early psychotherapy research was plagued by the fact that therapists subscribing to a particular practice method did not use consistent or distinctive techniques. Replication studies were difficult because descriptions of interventions were not sufficiently detailed to permit duplication. In later psychotherapy research, these problems were addressed through identification of critical ingredients and development of process rating scales for their measurement. Thus, establishing fidelity criteria and being able to measure adherence enabled treatments to be more standardized, consistently researched, and replicated.

Fidelity measures also provide methods to document deviations from an intended model and differences among the variations of a model (Bond, Becker, Drake, & Vogler, 1997). For meta-analyses, having measures of fidelity can assist in producing meaningful comparisons of treatments (Banks, McHugo, Williams, Drake, & Shinn, 2001; Bond, Williams, Evans, et al., 2000). Applying fidelity criteria in randomized clinical trials can assure that the experimental treatment is really absent in the control condition (Mills & Ragan, 2000). Program drift is common in community settings (Bond, Williams, et al., 2000), and the use of fidelity measures on an ongoing basis can warn that it is occurring. In multi-site studies, fidelity criteria are essential to ensuring that the services being studied are the same across sites, or at least that significant differences are documented and measured (Paulson, Post, Herincks, & Risser; 2002). If necessary, fidelity criteria may provide a basis for excluding data from sites which deviate too far from the experimental treatment model (Teague, Drake, & Ackerson, 1995). Further, when established models are replicated using valid criteria, measures of fidelity do predict outcomes (Blakely et al., 1987; Paulson et al., 2002). Model developers have noted that when key elements are left out of replications, less positive or even contradictory outcomes have been the result (Bond, Evans, et al., 2000). Finally, in research applications, well-developed and valid measures of fidelity can actually enhance statistical power in treatment outcome studies, acting as moderating variables to help explain variance in outcomes (Teague et al., 1995).

In terms of administrative issues, Unrau (2001) cites the Government Performance and Results Act of 1993 as motivating a need to describe and evaluate program delivery. Without specific criteria governing program operations, an innovative and non-traditional approach to service delivery (e.g., the wrap-around model in children's services) or an evidence-based practice can revert to merely the status quo in replications (Bilsker & Goldner, 2002; Malysiak, Duchnowski, Black, & Greeson, 1996). Thus, fidelity criteria can be used as a guide to implementing a program model as intended (Bond et al., 1997), or to monitoring programs to help assure quality (see e.g., Bond, Williams, et al., 2000). Having fidelity criteria should also promote external validity by providing adequate documentation and guidelines for replication projects adopting a given model. Several authors (e.g., Bond, Evans, et al., 2000; Brekke, 1988) have noted that fidelity criteria may be especially needed in the mental health field,

as programs often lack model specification and/or model adherence and rely extensively on clinical knowledge and skill.

Steps in Establishing Fidelity Criteria

Researchers (McGrew et al., 1994; Teague, Bond, & Drake, 1998) have described three major steps in establishing fidelity criteria. The first is to identify possible indicators or critical components of a given model (often based on an expert consensus process or the existence of a proven model which has been explicitly described), describing sources of data for each indicator, developing operational definitions for the indicators or critical components, including specifying anchors for points on rating scales, so that they are objective and measurable. The second step is to collect data to measure the indicators (preferably through a multi-method, multi-informant approach). The third step is to examine the indicators in terms of their reliability as well as validity (predictive, discriminant, construct, etc.) (Moncher & Prinz, 1991). Bond, Williams, et al. (2000) expanded the number of steps to 14, by providing a description of recommended activities before undertaking fidelity assessment and additional details on the process of measurement development.

LITERATURE REVIEW OF STUDIES ON FIDELITY ASSESSMENT

In the following sections of this paper, we review published studies that have developed, measured, validated, and/or used fidelity criteria and measures. The literature review is based on these bibliographic databases: Psych Abstracts, ERIC, Social Science Index, Social Work Abstracts, and MedSearch, focused on years 1995 up to the present. The literature encompasses services in mental health, health, substance abuse treatment, education, and social services. The review is organized according to the three major steps in establishing fidelity criteria—summarizing typical as well as more original approaches, providing examples, and identifying some major issues which are significant from either a scientific, logistical, or cost point of view. Table 1 provides information on fidelity studies: (1) how criteria were identified and developed, (2) how they were measured, (3) methods to assess reliability and/or validity, and (4) resulting scales or instruments. Published studies were included in the table if they provided information on at least two of the first three topics. Some studies are included in the narrative but not in the table, such as studies which described only the validation of fidelity criteria (e.g., Henggeler, Melton, Brondino, Scherer, & Hanley, 1997; Johnsen et al., 1999).

Step One: Identifying and Specifying Fidelity Criteria

Fidelity criteria should include aspects of structure and process (two of the three components of quality in the classic Donabedian, 1982, model). *Structure* encompasses the framework for service delivery, and *process* comprises the way in which services are delivered. Fidelity criteria often include: specification of the length, intensity, and duration of the service (or dosage); content, procedures, and activities over the length of the service; roles, qualifications, and activities of staff; and inclusion/exclusion characteristics for the target service population (Kelly, Heckman, Stevenson, & Williams, 2000).

Because the first step of identifying and specifying the criteria is the building block for assessing fidelity, one might think it should get the most attention. However, the articles

TABLE 1.

Article	Focus	How Criteria were Developed	How Criteria were Measured	How Criteria were Validated	Instrument Produced
Becker et al. (2001)	Supported employment (SE) model for adults with serious mental illness–Individual Placement & Support (IPS)	From IPS manual, authors' experience in implementing model & SE literature	Semi-structured interview (up to 1 hr) w-knowledgeable staff worker from program	Supp. employment progrs in 10 MH Centers rated on fidelity; 2 components correlated sig. with competitive empl. outcomes	
Blakely et al. (1987)		Intrvws & in-person observs of models and replication plus info published by developer – analyzed to delineate components as well as variations	Research staff-pair rated programs on fidelity scale, based on site visits and records	% exact agree. between raters; convergent validity–exact agree. between information sources; sig. correl. between fidelity score and outcome effectiveness	List of components ranged from 60 to 100 for each model program; rated as ideal, acceptable, unaccept.
Bond et al. (1997)	IPS (see Becker et al., 2001)	From IPS manual, authors' experience in implementing model & SE literature	Semi-structured interview (up to 1 hr) w-knowledgeable staff worker from program	Inter-rater & internal consist. reliability; IPS differentiated from other SE programs & from non-SE voc rehab programs	IPS Fidelity Scale, 15-items, 5-point ratings; 5 = ideal to 1 = contrary to standards
Clarke (1998)	Adaptation of Coping with Depression course for adolescents–prevention & treatment	Based on compliance with an existing treatment protocol	Sessions (live or on videotape) were rated on a fidelity scale by a supervisor or res.asst.; ratings were summed	Inter-rater & internal consistency reliability; too few groups to relate fidelity to outcomes in a prevention RCT trial	Fidelity scale, 10-items, 3-point ratings; 0 = no adherence to 2 = complete adherence
Friesen et al. (2002)	HEAD START and other early childhood programs	Qualitative study of 3 contrasting Head Start programs, plus lit review to develop conceptual framework and from this a scale	Survey of sample of personnel in Head Start progrs plus annual Program Information Reports	Relationship between survey results and these proposed DV's: % children referred for mh probs; % children receiving treatment	Under development
Hernandez et al. (2001)	Systems of care for families with SED child	Not clear. Used system of care values and principles which apparently evolved over time	Document reviews and interviews w-families by a team of 6 professionals trained in use of instrument	Examined scores for exemplary programs (top quartile) vs. traditional programs and found significant differences	System of Care Practice Review (SOCPR), 34 questions, 7-point ratings

Table 1 (*Continued*)

Article	Focus	How Criteria were Developed	How Criteria were Measured	How Criteria were Validated	Instrument Produced
Henggeler et al. (2002)	Family based mental health treatment–Multi-Systemic Therapy (MST)	Measure developed by expert consensus, and based on MST manual	Ratings of therapist adherence from phone interviews of caregivers once/mo., also youth ratings and therapist ratings	CFA, factor anal., test-retest correls., Cronbach alpha, correls. of supervisor/ therapist ratings; rel. of adherence to youth/family outcomes	Therapist Adherence Measure (TAM) and other MST adherence measures (26 items)
Kelly et al. (2000)	HIV prevention/ intervention programs funded by CDC	Core elements of intervention determined from participant feedback, experienced facilitators, and community advisors	Not Specified	Core elements should consistently relate to outcomes across sites and key characteristics may relate to outcomes at some sites	None
Lucca (2000)	Clubhouse model of vocational rehabilitation (VR) for adults with psychiatric disabilities	Reviewed mission statements and documents from selected clubhouses and published literature	22 programs; single informant at each program indicated presence/ absence of each index item (component)	Internal consistency reliability; sig. diffs for clubhouse vs. other VR models; sig. correl. btwn index score and Prins. of PSR scale	15-item index of components which should and should not be part of the model; marked yes/no
Macias et al. (2001)	Clubhouse model, based on Fountain House	Content analysis of ICCD certific. reports which used Clubhouse Standards. TF of clubhouse staff picked standards which discriminated between cert. x non-cert. clubhouses	Mail survey to program admins. in 166 clubhouses which had gone through the certification process.	Discriminant validity: Certified clubhouses endorsed sig. more items than noncertified. However, some items showed uniformity of responses	Clubhouse Research & Eval Screening Survey (CRESS) has 59 yes/no items, attempts to avoid subjective assessments
Malysiak et al. (1996)	Wrap-around model to provide mental health and case management services to children & adolescents w-emotional/behavioral disorders	Value-based philosophical principles; participatory evaluation involving program staff to describe what worked and what didn't work	Observation of team meetings, meetings with families and review of case files	No information	None

McGrew et al. (1994)	Adult mental health program–Assertive Community Treatment (ACT)	Interviews of ACT researchers and original program developers–asked importance of ACT critical components from published descriptions. Scale of fidelity resulted; expert judgements used to weight items. Scoring criteria operationalized 3 levels per item	Researchers reviewed write-ups and records of ACT programs, augmented by reports by program directors, site visitors, and consultants	Inter-item reliability; relationship between program fidelity score and program impact (# days hospitalized); fidelity scores for ACT vs. traditional case management.	Index of Fidelity for ACT (IFACT)–14 items
Mills & Ragan (2000)	Integrated Leaning Systems (type of computer technology used in educational software)	Telephone interviews of innovation developers to identify essential features; focus group of teachers who are users; construct a component checklist and pilot test.	Teacher completes checklist, teacher int. by researcher, observation of software in use. Panel of 3 experts–review transcriptions and independently score components	Scores were cluster analyzed; configuration patterns examined for differences–a number were significant	Integrated Learning System Configuration Matrix (ILSCM)–15 implementation components, each with 5 levels of variation
Orwin (2000)	Substance abuse services–multi site study	Expert panel generated list of 39 distinct services to be reported and glossary of terms providing common definitions, plus identifying dimensions for codifying programs v/v each activity	Participants reported whether they received service. Count up # services that were planned as part of model	Sites with multiple intervention conditions, and participants in more intensive groups more likely to get planned services	N/A
Paulson et al. (2002)	Consumer choice as a component of mental health/rehabilitation programs	Consumer consultants added questions re-choice making opportunities to an existing fidelity scale	External reviewers examined program documents, etc. and did ratings on criteria	Not yet validated	IPS+ – 41 questions covering 6 dimensions
Rog & Randolph (2002)	Supported housing, multi-site study	Steering Committee specified fidelity framework from RFA; defined major components and identified measurement indicators	Interviews with program management and staff, but not clear how this data were turned into fidelity scores	Comparison of supported housing vs. comparison programs v/v distance from ideal supported housing type	Fidelity instrument, not clear how many items or how they were scored
Teague et al. (1995)	ACT teams for mental illness/ substance abuse treatment (CTT)	9 ACT criteria from previous research, modified for the setting; 4 criteria on MI/SA added, based on researchers' experiences	Staff activity logs, agency docs. & MIS, site visits & intvs. – reviewed by research team to produce consensus ratings	7 CTT vs. 7 standard case mgt progrms compared; cluster analysis used to group sites	13 criteria, scored from 1-5 in half-point steps

Table 1 (*Continued*)

Article	Focus	How Criteria were Developed	How Criteria were Measured	How Criteria were Validated	Instrument Produced
Teague et al. (1998)	ACT teams	ACT criteria from previous research and published literature	Progr rpts. from supvs. or staff, agency documents, MIS, struct. intvs. w-multiple informants–reviewed by informed raters	Factor anal. and intern. consistency reliability; validation used 50 progrms differing in degree of intended replication of ACT	DACTS – 28 criteria, 5 point ratings
Umrau et al. (2001)	Family literacy program	1 day workshop–Comm. Stakeholders & program staff–produced program philosophy, goals, logic model & activities. Exit interviews w-families to identify pathways thru which outcomes were achieved	Daily activity checklists completed by workers	N/A	N/A
Vincent et al. (2000)	Pregnancy prevention program	Based on experiences in operating the original model in another state	Records & reports from original project, subjective perceptions of model developer; compared to replication site records, reports, exit interviews & community surveys. Researchers judged comparability between projects	N/A	N/A
Weisman et al. (2002)	Family Focused Treatment (FFT) for bipolar patients and their relatives	Scale based on treatment manual	Ratings from videotaped treatment sessions by 3 professionals trained in FFT	Inter-rater agreement (ICC's from .74-.98); rel. between fidelity score and patient outcomes (relapsed or not) NS	Therapist Competence/ Adherence Scale (TCAS)–13 items, 7-point scale

we reviewed often lacked detail about how their fidelity criteria were derived. For example, Hernandez et al. (2001) developed a measure of fidelity to system-of-care principles (in services to children with serious emotional disturbances). The domains and subdomains for the fidelity criteria are described, but no references are provided for their sources. Nor is there indication of how the specific principles were selected. Holden et al. (2002) evaluated a demonstration project of a community-based, continuum of care approach for children in the child welfare system, eligible for residential services. They discuss the importance of examining implementation fidelity and present operating principles, but not their origin or how their measures reflect the principles. Rog and Randolph (2002) present fidelity criteria and indicators for a supported housing multi-site study, which were developed by a Steering Committee, based on the RFA for the study; however, further details about the process, the criteria, or the indicators are not provided. Teague et al. (1995, 1998) drew on an earlier fidelity framework that had been derived from semi-structured interviews with Assertive Community Treatment (ACT) experts, rating key ingredients and specifying ideal levels of ingredients (McGrew & Bond, 1995). This earlier work had also shown correlations between some fidelity variables and outcomes (McGrew et al., 1994), but Teague et al. (1995) went beyond the limited available empirical evidence to add and operationalize criteria on the basis of the authors' appraisals of desirable program operations and feasibility of measurement.

Methods to Develop Fidelity Criteria

When the process of developing fidelity criteria is described, it has primarily involved one of three methods: (1) drawing from a specific program model with proven efficacy, effectiveness, or at least, acceptance; (2) gathering expert opinions—surveys of experts, and/or literature reviews; or (3) qualitative research—opinions of users and advocates regarding what works, site visits to diverse programs, etc. In the mental health literature, most of the current examples of fidelity measures have come from (1), the use of a specific program model that has some evidence of positive outcomes. An example is Fountain House, established over 50 years ago and now the model for a "clubhouse" approach to psychiatric rehabilitation. Clubhouses use the "work-ordered day" in which members (clients) work voluntarily in units necessary to run the clubhouse for the benefit of all the members. Thus members gain experience and skills in "real" not "make-work" jobs, which they are then enabled to use in transitional work outside the clubhouse. Fountain House developed structured training curricula and methods for certifying adherence to the model. Standards for clubhouse programs were then developed collaboratively by a set of founding clubhouses. Based on their own experiences and content analyses of certification assessments, a Task Force developed a screening tool for quantifying clubhouse fidelity, focusing on the standards most likely to discriminate between certified and non-certified clubhouses. The result is the Clubhouse Research and Evaluation Screening Survey, or CRESS (Macias, Propst, Rodican, & Boyd, 2001). A similar, but less detailed, example comes from Clarke (1998), who conducted trials to assess the effectiveness of an accepted intervention for depression with adolescents; fidelity criteria were based on an established treatment protocol and program manual.

Other fidelity measures have been based on models with established efficacy. For example, Drake, McHugo, and Becker (1996) developed a vocational rehabilitation program model for adults with psychiatric disabilities, Individual Placement and Support (IPS), in a federally-funded research project. The program produced significant positive findings in a randomized clinical trial. As part of this research, the program developers produced fidelity

criteria, a scale to measure fidelity, and a program manual (Bond et al., 1997). Similarly, Henggeler and Schoenwald (1998) developed a family-based therapy for juveniles in corrections and/or mental health treatment, Multisystemic Therapy (MST), and established its efficacy in clinical trials, later producing a program manual and fidelity scales, using expert consensus, based on the manual. With Family-Focused Treatment (FFT) for bipolar patients and their family members. Weisman, Tompson, Okazaki, and Gregory (2002) developed a scale to assess clinicians' fidelity to the program model, based on a treatment manual on FFT developed earlier. Blakely et al. (1987), through site visits, interviews, and reviews of written materials, identified components of the original models (from education and criminal justice) on which replications were built. A program component was defined as an activity, material, or facility which could be observed or verified, was logically discrete from other components, and was specific to the innovative program.

The first and most well-known program to develop fidelity criteria in the mental health field (outside of psychotherapy rating scales) is Assertive Community Treatment—generally recognized as the most widely tested and successful model of community-based treatment and rehabilitation for adults with serious mental illness (Mueser et al., 1998; Stein & Santos, 1998). However, in ACT, compared to more recent evidence-based models, the development of fidelity criteria (Teague et al., 1998) and the program manual (Allness & Knoedler, 1998) occurred much later (1994–1998) than the original efficacy study (Stein & Test, 1980). The first scale developed to assess fidelity to ACT principles, in fact, followed the expert opinion method for fidelity development approach (2) above: reviewing published descriptions of the model, constructing a list of proposed critical ingredients, then having ACT experts (academics and practitioners) rate the importance of each ingredient (McGrew et al., 1994). Subsequent ACT fidelity studies have built on these criteria, adjusting for specific settings and/or populations (Johnsen et al., 1999; Teague et al., 1995) and revising on the basis of new literature and measurement practicality (Teague et al., 1998). This general method has also been used to identify fidelity criteria for consumer-operated programs (Holter, Mowbray, Bellamy, & MacFarlane, in press). That is, published articles on these programs and the philosophy of peer-provided services were reviewed and criteria developed; the criteria were then rated by consumerism experts (consumers and non-consumers), using a modified Delphi method.

Early efforts to establish fidelity criteria for some models utilized only literature reviews—identifying "active" or "essential" ingredients (case management, Rapp, 1998; psychiatric rehabilitation, Anthony, Cohen, & Farkas, 1982); or a framework to measure implementation (Brekke & Test, 1992). However, in these efforts, the criteria were not operationalized nor developed into quantitative measures which could be analyzed, verified, and related to client outcomes or other indicators of impact.

Experts, but not literature reviews, were used in three other fidelity studies. Paulson et al. (2002) employed two consumer consultants to develop items on consumer choice to be added to the Individual Placement and Support fidelity scale. Orwin (2000), in a multi-site research demonstration project on alcohol treatment for homeless persons, utilized an expert panel which identified 39 distinct services which were to be reported. Unique to this study, a "leakage" scale was also developed, to measure programs delivering non-planned services. Maximum adherence to the intervention model was represented by a high score on fidelity and a low score on leakage. Vincent et al. (2000) based their assessment of the fidelity of a replication on historical records from the original model project plus the subjective judgments of the model originator.

Other fidelity studies have incorporated qualitative methods to identify criteria or critical components, often combined with other methods. To guide development of Head Start and other early childhood programs, knowledge obtained from published effectiveness studies was used to produce a conceptual framework, followed by qualitative research on three Head Start programs, selected because of their differing approaches to the delivery of mental health services (Friesen et al., 2002). The conceptual framework encompassed organizational variables (program size, auspices, staffing level and expertise), program philosophy (values, beliefs, ideologies), and program resources. Others (Giesler & Hodge, 1998; Kelly et al., 2000; Malysiak et al., 1996) used feedback from participants, staff, or other stakeholders to identify elements of program models that seemed to contribute most to their success. Unrau (2001) also did this for a family literacy program, by convening a one-day participatory workshop involving 20 different stakeholders, combined with exit interviews with a sample of families. Mills and Ragan (2000) probably used the most extensive qualitative methods in assembling fidelity criteria (for a computer-based, Integrated Learning System, ILS). They reviewed educational software vendor documentation and published research on ILS. They also conducted interviews with individuals who had developed, sold, or used the courseware and they convened a group of teachers who used the courseware, conducting focus groups, individual interviews, and observations of the courseware in use to produce a checklist of primary components. Lucca (2000) also used mixed methods to produce a list of service components for a fidelity measure of psychiatric clubhouse programs—including components that should and should not be part of the model. The information was derived from published literature about clubhouses, as well as reviews of mission statements and other documents from existing clubhouse programs. Finally, Blakely et al. (1987) based identification of critical ingredients on existing exemplary models, and, to specify scale points for the degree to which criteria were implemented, they visited program replicas to observe and document variability.

Thus, there is great diversity in methods to identify and specify fidelity criteria. This is less of a problem for programs which start as research/demonstrations, where fidelity criteria and a program manual are in place when the intervention begins. However, there are some studies utilizing fidelity criteria which have failed to adequately explain how the fidelity criteria were derived. It appears that the conceptual activity around developing fidelity criteria needs much more attention. We summarize the major issues in the next section.

Issues in Establishing Fidelity Criteria

From conceptual and logistical perspectives, probably the most appropriate and feasible method of establishing fidelity criteria is to examine the operations of a program model that has proven successful (Bond, Williams, et al., 2000). In more recent years, especially when funded by federal sources such as NIH, research designs are expected to have identified key components which are then tested rigorously through an efficacy study. Having a program manual in place to guide implementation is really a prerequisite to the receipt of funding for a randomized clinical trial; research funding is too scarce and RCTs too expensive and time-consuming to risk on a research design that does not guarantee implementation according to the model being tested. However, while this is the situation for many federally-funded efficacy studies (NIH, SAMHSA), usually models are not tested in such a thorough and comprehensive manner. Other funding sources may not have the same standards of rigor for efficacy or effectiveness research. A major challenge is the large number of programs already operating that have not been and are unlikely to be subjected to efficacy research. In those cases, it is still desirable to conduct high

quality program evaluations and to gather as much evidence regarding outcomes as possible. But to do such evaluations, we still need fidelity criteria to achieve some standardization about what it is we are studying; otherwise, the situation is similar to that of the early psychotherapy researchers—unsure about what it was they were evaluating. So how do we put together fidelity criteria for existing services?

Several of the articles reviewed described being confronted with just this challenge. In these cases, authors identified and consulted sources of expertise, including literature reviews of published research studies or panels of academic or service experts running these programs. However, for such approaches, the unspecified nature of the program model under study still constitutes a major issue in specifying fidelity. Malysiak et al. (1996) state that difficulties determining treatment fidelity for wrap-around services in children's mental health systems are due in large part to lack of an articulated theory. Bond, Evans, et al. (2000) note that the development of fidelity measures is hampered by the lack of well-defined models and that the identification of fidelity criteria and development of fidelity scales for ACT were so successful because this model was well-developed and its operations were specified in detail. Development of fidelity criteria is more difficult with complex interventions that depend on practitioner decision-making, using clinical expertise, on individualizing services to meet the multiple needs and preferences of consumers, or on behaviors of multiple practitioners, structural variables, or service coordination (Bond, Evans, et al., 2000; Teague et al., 1998).

Orwin (2000) suggests that before developing fidelity criteria, administrators and evaluators should first do an evaluability assessment—a front-end evaluation that enables managers to determine the extent to which the program can be appropriately evaluated, based on a detailed program description, for example. An evaluability assessment can thus help to identify poorly defined interventions and vague, unrealistic objectives. Bond, Evans, et al. (2000) commend the value of focusing on "prohibited behaviors"—actions or program characteristics which vary markedly from the program model and are thus to be avoided. Several of the studies reviewed did this. Lucca (2000) included in her checklist program services that clubhouses were not supposed to provide. In the CRESS instrument, one-fourth of the items were prohibitions (Macias et al., 2001). McGrew et al. (1994) included "distractor" items among standard ACT criteria. Orwin (2000) measured "leakage"—activities intended to be done at another site.

Where "expert" consensus is used to develop criteria, several issues arise. Evaluators should be aware that, in the absence of established empirical findings, opinions of experts may change significantly (sometimes appropriately) over time, and the predictive utility of expert opinion may be quite low. Schemes for grading levels of evidence for interventions place expert opinion lowest in the hierarchy of knowledge or discount it altogether (Centre for Evidence-Based Medicine, 2002; Chambless & Ollendick, 2001). Still, where a proven model is not available and the research base is limited, expert opinion may be the only, if provisional, alternative.

Another issue concerns the availability of such experts and their credibility (i.e., expertise according to whom?). In some newly developed models, such as consumer-operated services, there is little in the way of published literature, and the articles that do exist usually only describe programs, providing little if any evidence that the program described is a high quality or effective one. Similarly, there may be few established experts in a newly developing program area; in some of the studies described, the only experts were people who had run or received services from the program at local levels. There is also the need to recognize that there are different perspectives on what constitutes expertise. Consumer-providers might feel that academic researchers could never be experts on consumer-operated services, unless they were

consumers themselves, no matter how many programs they had evaluated. Bond, Williams, et al. (2000) recommend including multiple perspectives in expert review panels. But how many experts are enough? In the literature reviewed, the numbers varied from 2 (Paulson et al., 2002) to over 20 (Holter et al., in press; McGrew et al., 1994).

Another issue is the tendency for experts to rate the majority of components or criteria as "very important" (Bond, Williams, et al., 2000; Holter et al., in press). For this reason, forced rating methods or rank ordering of items may be preferred. Finally, while there may be general agreement about fidelity criteria, there is often less agreement about operational definitions of critical ingredients. As an example, McGrew et al. noted in 1994 that while experts agreed that ACT should use a team-based approach, there was at that time little consensus about what should constitute a team.

Developers of fidelity criteria also need to be aware of the fact that fidelity to program standards can be confounded with the competence of the program implementers (Clarke, 1998); skillful practitioners may implement intervention models better and achieve superior results (Luborsky, McLellan, Diguer, Woody, & Seligman, 1997). In order to tease out this phenomenon, fidelity criteria could include items concerning such features as expected staff experience and training and monitoring of staff performance. Developers should also consider the dynamic nature of programs and service delivery. Programs usually need to undergo periodic changes in response to changes in client needs and the context of other resources available. Are the fidelity criteria developed today likely to be indicators of programs that achieve success indefinitely into the future? Given rapidly changing political, economic, and clinical circumstances, that seems unlikely. As an example, some case management programs have incorporated various features of ACT, resulting in much smaller measurement differences in fidelity between ACT and traditional programs (Teague et al., 1998). Do we need to raise the bar? If program effectiveness is a dynamic construct and programs need to periodically adjust to remain effective, how often do we need to re-examine fidelity criteria and consider their revision?

This relates to a major issue in the dissemination and implementation literature—that of adaptation. Historically, there has been a tension, even debate, between schools of thought that advocate for exact replications of effective program models (Drake et al., 2001; Szulanski & Winter, 2002) versus the need to adapt models to local conditions to maximize efficiency as well as local ownership (Bachrach, 1988; Fairweather & Tornatzsky, 1971). Adaptation may be necessary due to special needs of the target population, differences in budget, community resources, or organizational factors (Johnsen et al., 1999). On the other hand, it is generally agreed that programs with higher fidelity to efficacious models produce superior outcomes (Blakely et al., 1987; Drake et al., 2001). Determining which components of the program are essential, irrespective of context, and therefore require absolute fidelity to the original model, and which components may be modified, eliminated, or added, is an empirical matter. This point is further elaborated in the section on criteria validation and addressed in the discussion.

Step Two: Measuring Fidelity

Published articles have generally provided detailed information about this second step—developing and implementing methods to measure (quantify) adherence to fidelity criteria. The most common methods to quantify fidelity are: (1) ratings by experts, based on project documentation and/or client records, site observations, interviews, and/or videotaped sessions;

and (2) surveys or interviews completed by individuals delivering the services or receiving them.

Mills and Ragan (2000) provide one of the most detailed and clearest examples of ratings by experts to measure fidelity—a study of Integrated Learning Systems. Teachers who used the software under study completed checklists (based on fidelity criteria). One of the researchers conducted 45-minute, semi-structured interviews with these teachers, which were tape-recorded. A panel of three experts reviewed all transcribed audiotapes and independently scored each response on the fidelity measure. Each program's cumulative raw scores were then standardized. A similar, comprehensive, multi-source method was used by Hernandez et al. (2001) to rate services for children using the System of Care Practice Review (SOCPR). A team of 5–6 individuals was trained for three full days on the SOCPR. Administration of the protocol was structured into sections: record-keeping instruments, reviews of treatment plans and individualized educational plans from case records, and interviews of caregivers, children/youth, and providers. For the last section, reviewers provided ratings (based on information summarized in the other sections) on 34 summative questions.

Also using expert raters, Johnsen et al. (1999) used the DACTS (Teague et al., 1998) to measure fidelity to the ACT model in 18 case management programs for persons with mental illnesses who were homeless. Raters used information from each program's application for continuation funding, annual site visits which included interviews with staff and administrators, and feedback from program directors in response to tabular summaries of the above. In other assessments of ACT, information was compiled from clinicians' activity logs, agency documents, management information systems (MIS), site visits, and informal interviews for the research team to review and come to consensus ratings (Teague et al., 1995, 1998). Similarly, Malysiak et al. (1996) assigned programs to model categories based on three primary sources of data: observation of team meetings, interviews with some participants, and review of case files. To assess the replication of a community-based, teen pregnancy prevention model, Vincent et al. (2000) compared historical records and the perceptions of the model developer concerning the original model to documentation on the replication, in order to produce a narrative summary of similarities and differences. None of these three studies detailed who the raters were or whether or how they were trained.

In other studies, the information gathering of raters is far less extensive. For example, completion of the fidelity scale or program component checklist has been based on interviews of knowledgeable staff (Bond et al., 1997), in some cases supplemented by site visits to programs (Becker, Smith, Tanzman, Drake, & Tremblay, 2001; Blakely et al., 1987). Ratings of Assertive Community Treatment Programs, using the IF-ACT (McGrew et al., 1994), were obtained in one-hour, semi-structured interviews with each program director, supplemented by reports from site visitors and consultants. The rating of fidelity to a Family-Focused Treatment model involved three experts trained in FFT, utilizing the Therapist Competence/Adherence Scale, after viewing the videotape of the first family session in each segment of treatment. For rating sessions of the intervention in the Coping with Depression Course-Adolescents, raters used a project-developed fidelity scale (Clarke, 1998).

Examples of direct data collection from staff or participants to measure fidelity are fewer and quite diverse in their methods. Friesen et al. (2002) are using a measure of organizational philosophy as one part of fidelity assessment, obtained through interviews with Head Start administrators, management staff, teachers, and parents. In the Henggeler, Schoenwald, Liao, Letourneau, and Edwards (2002) study of fidelity, a Therapist Adherence Measure was administered to families receiving MST services, after the start of treatment and monthly there-

after, through phone interviews by an MST employee other than the family therapist. Paulson et al. (2002), examining extent of consumer choice in an employment intervention, utilized a structured fidelity scale completed by staff and consumers at program start up and every six months. Unrau's (2001) measure of fidelity comes from daily activity checklists completed by workers after each family session. Orwin (2000) used Quarterly Report Forms, completed by site personnel, to produce each site's fidelity and leakage scores, based on reports of types of services provided. Lucca (2000) had a single staff person complete a checklist of program service components proposed as essential to clubhouse operations. The CRESS (Macias et al., 2001), which also measured fidelity for clubhouse programs, was administered through a mail survey, to be completed by program administrators. The survey contains 64 questions which are either dichotomous items, composite items (subsets of yes/no contingency questions) or checklists—all of which have a single correct response.

In short, detailed descriptions of the measurement of program adherence to fidelity criteria are often included in published reports. The diversity of methods and sources is noteworthy and should be of assistance to other program evaluators, seeking to utilize multiple methods and multiple sources to establish fidelity—a recommended practice. The methods presented, however, do present several issues which need to be addressed.

Issues in Measuring Fidelity

Some measurement issues are those typically found in any field research study, involving the reliability and validity of the measurement devices. For example, relying on therapists or other staff to accurately report their activity (or lack thereof) may limit actual or perceived validity, through a social desirability bias, especially if staff suspect that the ratings may affect program funding. Asking service users to provide ratings usually involves individuals who volunteer to do so. However, research suggests that volunteer participants are oftentimes biased in terms of being overly positive or overly negative about the evaluand (Lebow, 1983; Nguyen, Atkisson, & Stegner, 1983). Even in a representative sample, factors beyond the program characteristics themselves are known to affect the variability of such assessments (Teague, 2000). Unrau (2001) advises that measurement of fidelity should not use client data alone, but rather, that client data should complement other evaluation approaches. Using researchers to produce ratings across program types may also pose validity problems, in that usually there is no way to keep them blind to the type of program they are rating. These issues are lessened when the fidelity scale utilizes objective, behaviorally anchored criteria for each scale point, involving little inference (Bond et al., 1997). The CRESS measure of fidelity to clubhouse standards used dichotomous items rather than Likert rating scales to minimize subjective assessments (Macias et al., 2001).

Some researchers have noted that not all fidelity criteria are measurable with the same reliability, feasibility, or cost. Oftentimes the distinction is made between structure and process criteria. Structure encompasses measures of staffing levels and characteristics, case load size, budget, procedure codes, frequency and intensity of contacts, etc. (Orwin, 2000). Structure measures require less subjective judgment and can often be obtained through existing documentation. Process criteria include program style, staff–client interactions, client–client interactions, individualization of treatment, or emotional climate (e.g., hostility, chaos, organization, empowerment). Rating program performance in relation to these criteria requires more subjective judgments, often based on observations, interviews, and other data sources, and thus necessitates more time and effort, is much more costly, and is likely to be less reliable (Bond

et al., 1997), even if response scales are well anchored. McGrew et al. (1994) acknowledge that their IF-ACT contained 17 items that were readily measured and already available in existing program evaluations; nearly 60 other criteria endorsed by experts, including items in the domains of service coordination and treatment goals, were not measured. This resulted in a reliable scale, but uneven coverage of the program's operating constructs. Lucca's (2000) measure of fidelity to the clubhouse model included only a checklist of service dimensions. Process criteria may be more difficult to measure reliably, but more significant in terms of program effects. In studies of Assertive Community Treatment, model drift (from fidelity) occurred less on structural features and more on so-called discretionary (process) features, such as overall treatment approach and in vivo services (Teague et al., 1998).

The structure versus process debate relates to a larger issue, discussed in the first section, concerning how fidelity criteria are identified. Most approaches imply that their criteria were selected to reflect the model's most significant program components. However, even if the criteria presume to tap these components, there is little illusion that a practical fidelity instrument can measure them comprehensively. In many instances, the elements of a fidelity measure serve, in effect, as indicators of the model's design and operations—key program features that relate strongly to positive outcomes for those served—but do not necessarily include all such features, nor any features in the depth suggested by a fully explicated program theory. Indicators are selected, then, on an empirical basis (relationship with outcomes), and also because they are reliable and easy to measure. However, as programs and their contexts evolve, relationships with outcomes may change over time. To some degree this may occur because those indicators that are easy to measure may also be easy to manipulate so as to "game" the system. An example would be a program that improves its staff to client ratio and ensures that staff represent professionals with specified credentials, but still does not incorporate sufficient training or monitoring of staff–client interactions to ensure that staff do give clients choices and promote decision-making and empowerment. "Gaming" obviously reduces the validity of the fidelity measure. This phenomenon is often noted in evaluations in which performance indicators are used to make important decisions about continuance of a program, funding levels, staff salaries, etc. (Teague, Ganju, Hornik, Johnson, & McKinney, 1997). A well-designed measure, like a well-designed system of indicators, would anticipate this vulnerability and attempt to measure the more critical features as well, albeit at potentially greater effort. For example, it may be that, if program users are more active stakeholders in assessing fidelity, indicators of structural features may be more effectively complemented by indicators of critical processes.

This process issue involves the dynamic nature of programs: Programs are usually not static and often undergo substantial changes over time. But fidelity is frequently measured only at one point in time to answer the question of whether the replication is faithful to the original model or not (Bond, Evans, et al., 2000). However, the variables measured are not static traits. Investigators need to document and account for potential changes over time, especially if the fidelity measure is used for monitoring purposes. But how often should such measures be repeated? Too often unnecessarily consumes resources and results in unmanageable and uninterpretable data. Not often enough may mean losing an opportunity to meaningfully intervene before too much program drift occurs. A related measurement design issue is the question of how many sources to use in multi-method measurement approaches. Collection of data from multiple sources without careful examination of inevitable differences in perspective and response characteristics can complicate subsequent reconciliation and interpretation.

Step Three: Assessing the Reliability and Validity of Fidelity Criteria

The studies reviewed assessed reliability and validity using one or more of five different approaches:

1. Examining reliability across respondents, calculating the extent of inter-rater agreement (coefficient kappa, intra-class correlations [ICC], percent agreement, or Pearson correlations). For example, Henggeler et al. (2002) examined test–retest correlations of families' multiple ratings of their therapists, as well as the correlations between ratings done by therapists and by their supervisors. Weisman et al. (2002) reported the ICCs of the ratings by three professionals based on therapist videotapes. Clarke (1998) had a second rater for 14 sessions from the adolescent depression treatment study and calculated kappa statistics on fidelity assessments.
2. Examining the internal structure of the data empirically and in relationship to expected results, such as through confirmatory factor analysis (CFA), internal consistency reliability (Cronbach's coefficient alpha), or cluster analysis. For example, Bond et al. (1997), Clarke (1998), Lucca (2000), and McGrew et al. (1994) reported internal consistency reliabilities, while Henggeler et al. (2002) reported results of CFA and internal consistency. Mills and Ragan (2000) used cluster analysis to identify configuration patterns from teacher reports. That is, they tested each implementation component of their ILS software for significant differences in configuration patterns.
3. The method of known groups—examining differences in fidelity scores across types of programs expected to be different. Examples include comparisons between the top quartile of "exemplary" programs and traditional programs (Hernandez et al., 2001); ACT programs versus traditional case management (Teague et al., 1995, 1998); variations of supported employment programs (Bond et al., 1997) or clubhouses (Lucca, 2000) compared to traditional vocational rehabilitation models; supported housing demonstration sites versus comparison programs (Rog & Randolph, 2002); or sites with multiple intervention conditions relative to single interventions and participants in more compared to less intensive groups (Orwin, 2000). Bond et al. (1997) examined effect sizes in program comparisons and Teague et al. (1995, 1998) used cluster analysis to reveal program groupings.
4. Convergent validity—examining the agreement between two different sources of information about the program and its operations. For example, Blakely et al. (1987) compared records and documents with on-site observations. In a unique approach, Macias et al. (2001) examined self-ratings of compliance with clubhouse standards on the CRESS to the results from on-site, extensive certification procedures, comparing CRESS scores of certified to non-certified agencies. McGrew, Pescosolido, and Wright (2003) sought additional validation of ACT criteria by surveying ACT team members as to the extent to which they considered the critical activities involved to be beneficial. Lucca (2000) also examined correlations between Clubhouse fidelity index scores and scores on a Principles of Psychosocial Rehabilitation scale, to address convergent validity.
5. Examining the relationship between fidelity measures and expected outcomes for participants. The examples of this approach are: (a) fidelity scores for supported employment programs were related to client employment outcomes (Becker et al., 2001); (b) fidelity to the ACT model was significantly related to rates of hospital reduction (McGrew et al., 1994), and additionally; (c) for teams targeting co-occurring addictive

and mental disorders, to higher rates of retention in treatment, greater reduction in alcohol and drug use, and higher rates of remission from substance use disorders (McHugo, Drake, Teague, & Xie, 1999); (d) survey results from Head Start personnel were related to percentages of children referred for mental health problems and receiving mental health treatments (Friesen et al., 2002); and (e) fidelity scores on Family-Focused Therapy were examined for relationships with patient relapse (Weisman et al., 2000).

Issues in Validating Fidelity Measures

Two issues deserve attention. The first is the timing issue—that is, how to use fidelity measures on an ongoing basis. Several researchers have noted that, over time, the treatment as usual or comparison condition may become more like the program model (e.g., Teague et al., 1995). Fidelity measures may have to be redesigned, rescaled, or recalibrated to permit more sensitivity. The second issue concerns validating fidelity using client outcome measures. This seems problematic if fidelity is supposed to play a key role as a mediator or moderator variable in testing the effectiveness of the model. That is, if we have a model that is operating the way it is supposed to, will it produce the desired outcomes for clients? For example, Henggeler et al. (1997) examined the effectiveness of replications of Multisystemic Therapy in ordinary clinical practice and found that outcomes were better in cases with high treatment adherence. If a fidelity measure appropriately models a valid program theory, that is, if the measure is valid and reliable and the properly implemented intervention can actually produce the intended outcomes, then outcomes may be expected to vary with fidelity. Findings of such positive relationships serve as partial validation of both the program theory and the fidelity measure, with strength of demonstrated validity being a function of size and variability of the sample of programs evaluated. Under most circumstances, however, a sample of programs will be inadequate for these purposes. Thus, it seems desirable for validation purposes to examine fidelity measures for model replicas compared to other treatment programs serving the same populations and to test for significant differences, or to examine convergent validity (information about a single program, but obtained from differing sources, such as records, client or key informant reports, site visits for certification purposes).

Third, there is a level of analysis issue when comparing fidelity scores with data obtained from units within programs (such as clients in a program, or multiple records from a single program). Most of the analyses that have attempted to validate fidelity criteria have done so by simply aggregating individual data within programs and conducting analysis at the program level, ignoring within-program variability. Others have done analysis at the individual level, coding program-level variables as attributes of all units associated with the program, ignoring the fact that the individual units are not independent. Neither approach to analyzing nested data is optimally appropriate but may have been necessary in the past due to limited resources for analysis (Snijders & Bosker, 1999). However, with current statistical software and methods readily available (Hedeker, McMahon, Jason, & Salina, 1994; Raudenbush & Bryk, 2001), this is no longer the case. The level of analysis issue is discussed further in the next section.

DISCUSSION

We now turn to a discussion of some of the cross-cutting issues previously presented, as well as ideas for improving the methods used in development, measurement, and validation of fidelity

criteria. We believe that improvements could result in more widespread use of fidelity criteria for policy, practice, and evaluation purposes.

Development and Application of Fidelity Criteria

Fidelity criteria deserve more attention in research and evaluation studies. Their construction is not merely a technical exercise, but one which involves making many choices, especially when there is no existing program model with established efficacy on which fidelity criteria can be based. For example, how should fidelity criteria get developed? Who are the experts and how should they be involved? How are the fidelity criteria scaled and measured? Current approaches vary in the meaning of low fidelity scores (services as usual, total absence of services, or unacceptable services) (Teague et al., 1998). There may be a tendency to assign programs a single fidelity score, adding up ratings on separate criteria. However, managers and evaluators need to understand that, given the current state of fidelity measurement, a number of programs may all receive the same fidelity score, but be very different in their operations (Salyers et al., in press). Evaluators need to examine the structure of fidelity measures and consider presenting subscores on important but significantly different critical components.

In research, evaluators should more consistently consider measurement of fidelity criteria in control as well as in treatment conditions, allowing a comprehensive assessment of the extent to which both populations are receiving the "critical ingredients" of the program model. Bond, Williams, et al. (2000) note that "Careful psychometric work is needed if fidelity measures are to achieve their promise" (p. 18).

Another area of decision-making concerns the need to balance structure and process criteria. A focus on structural criteria may produce high reliability and validity at the cost of overly simplistic conceptions of program operations, while omitting key ingredients which are complex, reflecting values and principles, and which are, perhaps, more important (see Hermann et al., 2002, report on process measures of the quality of care for schizophrenia). Further, measurement development needs to address the dynamic nature of service provision—how do we assess and revise fidelity criteria over time? How do we validate differences? McGrew et al. (1994) found that ACT programs showed program drift over time—fidelity to some criteria decreased over generations of programs. But client populations also differ over time, so changes in fidelity to a program model or in a set of measures developed at another time or for another context may be appropriate.

There is also the related issue of basing fidelity criteria on easily measured indicators of program operations versus measuring the extent to which more subtle critical components are present and operating as expected. The dilemma is that variables based heavily on easily observable structural aspects are easier to develop, more timely, more reliable and, therefore, perhaps more appropriate in initial evaluations. However, in the long run, the predominantly structural approach is more subject to manipulation and more likely to need frequent revisions, updating, and empirical assessment. This may imply the desirability of formalizing a two-stage approach. When programs are initially being evaluated and fidelity criteria are first developed, an emphasis on structure over process items may be appropriate. Then if evaluations do show positive outcomes, measures of structure may permit more rapid initial replications of the program model, assuring fidelity at least to some significant components. However, for enduring programs and to facilitate movement toward more valid, mature replications, time should be invested in careful development and testing of reliable and valid measures of process criteria associated with critical components and based on program theory at a deeper level.

Validation of Fidelity Criteria

Attention to systematic and appropriate approaches to validating fidelity criteria are also needed. We suggest that for fidelity criteria to be useful in policy, practice, and effectiveness research, more is needed than analyses of their psychometric qualities, such as confirmatory factor analysis and internal consistency reliability. Validation studies must also go beyond construct and face validity. But rather than focusing primarily on outcomes as validators, fidelity criteria should be examined with regard to content and criterion validity. For example, programs that are intended to replicate a model having specified structure, methods, principles, and values, can be compared with programs that serve the same population but use distinctively different methods. Or, the same measures of fidelity can be compared across diverse information sources (staff, records, observations; Blakely et al., 1987) or against what a representative sample of program recipients say they are getting (Orwin, 2000). In our current, NIMH-funded study of consumer-run mental health programs, we are assessing comparable measures of the same fidelity criteria through off-site data collection (director interviews, staff questionnaires, record and document reviews), on-site data collection (structured staff observations of interpersonal interactions, interior space, exterior space, etc.), and structured interviews with the program participants concerning what they do at the program, how they interact with staff and other participants, and other measures of the criteria by which the programs are supposed to operate.

The statistical methods applied to fidelity data should be appropriate to the organizational levels at which the data are collected. Traditional analytic methods require that all data be aggregated or otherwise configured to a single specified "unit of analysis"—the program, in most fidelity research to date. Researchers have seen this as a problem when relating fidelity scores to outcomes, because a large number of programs had to be studied, in order to achieve the necessary power (Clarke, 1998). Newer multilevel or hierarchical approaches (cf., Hedeker et al., 1994; Raudenbush & Bryk, 2001) afford much greater flexibility, allowing analysis of relationships between program-level variables and characteristics of the individual participants that are nested within each program (e.g., individual-level outcomes or participants' reports about program operations). These methods appropriately model the dependencies inherent in data from participants in the same program while maximizing statistical power to identify cross-level relationships. They provide estimates of the magnitude of between-program variation relative to variation between participants in the same program, effectively providing a base rate for judging the impact of differences in program fidelity. Multilevel methods also allow for statistical adjustment for potentially confounding influences at the level of either program or participant, and make it possible to assess the extent to which the validity of fidelity criteria may be moderated by characteristics of either program (e.g., program tenure) or participant (e.g., length of program involvement). In our study of 30 consumer-run mental health programs, we are using multilevel methods to assess convergence between fidelity information gathered at the program level (e.g., documentation of consumer representation on the agency board of directors, observation of staff–participant interactions) and at the level of individual participants in each program (e.g., participant ratings of the extent to which consumers are involved in decision-making).

Fidelity—Enough or Too Much?

More research attention is also needed to issues of exact replication versus adaptation. There has been a long-standing controversy between maximizing fidelity in contrast to the

desirability or even the necessity of adaptation. Szulanski and Winter (2002) state that a best practice should be copied as closely as possible, in minute detail, and that adapting a successful template is a mistake. Drake et al. (2001) emphasize replication because fidelity ratings have been shown to relate to positive program outcomes. However, there is often a legitimate need to tailor a program model to local circumstances and resources and to the social and cultural needs of local participants (see Hohmann & Shear, 2002). For example, populations in different locations have different strengths and needs; service systems have different goals and objectives; and communities often differ widely in availability of and access to resources (Bachrach, 1988). Also, many studies show that local buy-in, as manifested in some program adaptation, is necessary in order to maintain ongoing program operations (Blakely et al., 1987). Clearly, the degree of adaptation observed or perceived as necessary by local implementers will also vary with the degree and level of program specification of the original model. But how do we determine which critical ingredients are essential to the model and the expected outcomes relative to those that may be adapted, omitted, or added at any given site? Several different approaches have been suggested. The empirical approach is to deconstruct program models and systematically test out the impact of key ingredients across sites. Then, elements that are found to be non-essential across the sites may be adapted, while elements that are really critical may not be changed (Kelly et al., 2000; Leff & Mulkern, 2002). Others have suggested a practical approach—that is, adaptation is acceptable up to a zone of drastic mutation (Hall cited in Blakely et al., 1987). This, of course, leaves a lot to variations in subjective judgments vis à vis what constitutes a "drastic" change. Finally, taking a theoretical approach, adaptations to local circumstances are seen as appropriate as long as they do not contradict the underlying program theory. In this approach, the program theory provides a "cognitive blueprint" for action (Price, Friedland, Choi, & Caplan, 1998). Staff are not expected to follow process protocols exactly, but rather, according to their own judgments of what fits with the client characteristics and context and the program theory. This approach is congruent with research about professionals being more engaged, motivated, and effective when they feel they are exercising their judgment and expertise (Glisson & Hemmelgarn, 1998; Henry, Butler, Strupp, & Schacht, 1993). Babor, Steinberg, McRee, Vendetti, and Carroll (2001) describe a successful multi-site evaluation (the Marijuana Treatment Project), which was manual-guided, not manual-driven. That is, therapists were asked to integrate the project's theoretical framework into their work, but not told to deliver therapy in a prescribed way.

According to Berman (1981), although there is no one best strategy, expecting perfect fidelity is more appropriate with well-specified innovations, whereas adaptive strategies are more effective with relatively unstructured innovations. Such an approach could help answer the question of how specific measurements of fidelity should be made (McGrew et al., 1994). Orwin's (2000) argument for flexibility is then more applicable to less structured innovations.

However, Berman's conclusion is still subject to debate, as some authorities assert that for an innovation to be successful, there must be mutual adaptation—a program model is adapted to local circumstances, but accompanied by changes in the organizational behavior of the adopting agency (Blakely et al., 1987; Price et al., 1998). The findings of Blakely et al. (1987) support this contention. In their study, while fidelity scores were significantly related to program effectiveness, so were measures of program additions—that is, modifications to the original model by the adopting agency. From our review, Blakely and colleagues appear to be the only investigators to research this important issue. It seems likely that needed resolution of the tension can best come from improvements in empirically supported advances in program

theory, so that each program criterion is specified at the optimal level, allowing choice between alternative pathways to fit local contexts.

CONCLUSIONS

In the foregoing article, we reviewed rationales, examples, methods, and challenges in measurement of program fidelity. We hope to have shown that replications of demonstrated models are appropriate and potentially fruitful objects of research, as are the processes of implementing a given model, documenting the basis of adaptation, and constructing and validating new measures or modifying existing ones to reflect fidelity to altered models. Teague et al. (1995) noted the need for implementation studies to examine organizational issues vis à vis their positive and negative contributions to model fidelity. This area is but one aspect of a host of contextual factors in both implementation and operation that may moderate the impact of a given program model (Hohmann & Shear, 2002). As progress is made in optimizing internal program specification, fidelity assessment may need to take more fully into account the role of context in future generations of measures.

ACKNOWLEDGMENTS

Supported, in part, through a grant from the National Institute of Mental Health to the University of Michigan, School of Work, R24-MH51363-06.

REFERENCES

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

Anthony, W. A., Cohen, M., & Farkas, M. (1982). A psychiatric rehabilitation treatment program: Can I recognize one if I see one? *Community Mental Health Journal*, *18*, 83–96.

Babor, T. F., Steinberg, K. L., McRee, B., Vendetti, J., & Carroll, K. M. (2001). Treating marijuana dependence in adults: A multi-site, randomized clinical trial. In J. M. Herrell & R. B. Straw (Eds.), *Conducting multiple site evaluations in real-world settings. New Directions for Evaluation, no. 94* (pp. 17–30). San Francisco, CA: Jossey-Bass.

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

Banks, S., McHugo, G. J., Williams, V., Drake, R. E., & Shinn, M. (2001). A prospective meta-analytic approach in a multi-site study of homelessness prevention. In J. M. Herrell & R. B. Straw (Eds.), *Conducting multiple site evaluations in real-world settings. New Directions for Evaluation, no. 94* (pp. 45–59). San Francisco, CA: Jossey-Bass.

Becker, D. R., Smith, J., Tanzman, B., Drake, R. E., & Tremblay, T. (2001). Fidelity of supported employment programs and employment outcomes. *Psychiatric Services*, *52*, 834–836.

Berman, P. (1981). Educational change: An implementation paradigm. In R. Lehming & M. Kane (Eds.), *Improving schools: Using what we know* (pp. 253–286). Thousand Oaks, CA: Sage.

Bilsker, D., & Goldner, E. M. (2002). Routine outcome measurement by mental health-care providers: Is it worth doing? *The Lancet*, *360*, 1689–1690.

Blakely, C. H., Mayer, J. P., Gottschalk, R. G., et al. (1987). The fidelity-adaptation debate: Implications for the implementation of public sector social programs. *American Journal of Community Psychology*, *15*, 253–268.

Bond, G. R., Becker, D. R., Drake, R. E., et al. (1997). A fidelity scale for the individual placement and support model of supported employment. *Rehabilitation Counseling Bulletin*, *40*, 265–284.

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

Bond, G. R., Williams, J., Evans, L., et al. (2000). *PN-44-psychiatric rehabilitation fidelity toolkit*. Cambridge, MA: Human Services Research Institute.

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., & 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.

Centre for Evidence-Based Medicine. (2002). *Levels of evidence and grades of recommendation*. Oxford, England (http://minerva.minervation.com/cebm/).

Chambless, D. L., & Ollendick, T. H. (2001). Empirically supported psychological interventions: Controversies and evidence. *Annual Review of Psychology*, *52*, 685–716.

Chen, H. (1990). *Theory-driven evaluations*. Thousand Oaks, CA: Sage.

Clarke, G. (1998). Intervention fidelity in the psychosocial prevention and treatment of adolescent depression. *Journal of Prevention and Intervention in the Community*, *17*, 19–33.

Donabedian, A. (1982). *The criteria and standards of quality*. Ann Arbor, MI: Health Administration Press.

Drake, R., Goldman, H., Leff, H., Lehman, A., Dixon, L., Mueser, K., & Torrey, W. (2001). Implementing evidence-based practices in routine mental health service settings. *Psychiatric Services*, *52*, 179–182.

Drake, R. E., McHugo, G. J., & Becker, D. R. (1996). The New Hampshire study of supported employment for people with severe mental illness. *Journal of Consulting and Clinical Psychology*, *64*, 391–399.

Fairweather, G. W., & Tornatzsky, L. G. (1971). *Experimental methods for social policy research*. New York, NY: Pergamon Press.

Frese, F. J., Stanley, J., Kress, K., & Vogel-Scibilia, S. (2001). Integrating evidence-based practices and the recovery model. *Psychiatric Services*, *52*, 1462–1468.

Friesen, B. J., Green, B. L., Kruzich, J. M., Simpson, J., et al. (2002). *Guidance for program design: Addressing the mental health needs of young children and their families in early childhood education settings*. Retrieved May 23, 2002, from Portland State University, Reseach & Training Center on Family Support and Children's Mental Health Web site: http://www.rtc.pdx.edu/pgProjGuidance.php.

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

Glisson, C., & Hemmelgarn, A. (1998). The effects of organizational climate and interorganizational coordination on the quality and outcomes of children's service systems. *Child Abuse and Neglect*, *22*, 401–421.

Hedeker, D., McMahon, S. D., Jason, L. A., & Salina, D. (1994). Analysis of clustered data in community psychology: With an example from a worksite smoking cessation project. *American Journal of Community Psychology*, *22*, 595–615.

Henggeler, S. W., Melton, G. B., Brondino, M. J., Scherer, D. G., & Hanley, J. H. (1997). Multisystemic therapy with violent and chronic juvenile offenders and their families: The role of treatment fidelity in successful dissemination. *Journal of Consulting and Clinical Psychology*, *65*, 821–833.

Henggeler, S. W., & Schoenwald, S. K. (1998). *The MST supervisory manual: Promoting quality assurance at the clinical level*. Charleston, SC: MST Institute.

Henggeler, S. W., Schoenwald, S. K., Liao, J. G., Letourneau, E. J., & Edwards, D. L. (2002). Transporting efficacious treatments to field settings: The link between supervisory practices and therapist fidelity

in MST programs. *Journal of Clinical Child and Adolescent Psychology*, *31*, 155–167.

Henry, W. P., Butler, S. F., Strupp, H. H., & Schacht, T. E. (1993). Effects of training in time-limited dynamic psychotherapy: Changes in therapist behavior. *Journal of Consulting and Clinical Psychology*, *61*, 434–440.

Hermann, R. C., Finnerty, M., Provost, S., Palmer, R. H., Chan, J., Lagodmos, G., Teller, T., & Myrhol, B. J. (2002). Process measures for the assessment and improvement of quality of care for schizophrenia. *Schizophrenia Bulletin*, *28*, 95–104.

Hernandez, M., Gomez, A., Lipien, L., Greenbaum, P. E., et al. (2001). Use of the system-of-care practice review in the national evaluation: Evaluating the fidelity of practice to system-of-care principles. *Journal of Emotional and Behavioral Disorders*, *9*, 43–52.

Hohmann, A. A., & Shear, M. K. (2002). Community-based intervention research: Coping with the "noise" of real life in study design. *American Journal of Psychiatry*, *159*, 201–207.

Holden, E. W., O'Connell, S. R., Connor, T., Branna, A. M., Foster, E. M., Blau, G., & Panciera, H. (2002). Evaluation of the Connecticut Title IV-E Waiver Program: Assessing the effectiveness, implementation fidelity, and cost/benefits of a continuum of care. *Children and Youth Services Review*, *24*, 409–430.

Holter, M. C., Mowbray, C. T., Bellamy, C., MacFarlane, P., & Dukarski, J. (in press). "Critical ingredients" of consumer run services: Results of a national survey. *Community Mental Health Journal*.

Institute of Medicine. (2001). *Improving the quality of long-term care*. Washington, DC: National Academy Press.

Johnsen, M., Samberg, L., Calsyn, R., Blasinsky, M., et al. (1999). Case management models for persons who are homeless and mentally ill: The ACCESS demonstration project. *Community Mental Health Journal*, *35*, 325–346.

Kelly, J. A., Heckman, T. G., Stevenson, L. Y., & Williams, P. N. (2000). Transfer of research-based HIV prevention interventions to community service providers: Fidelity and adaptation. *AIDS Education and Prevention*, *12*, 87–98.

Lebow, J. (1983). Research assessing consumer satisfaction with mental health treatment: A review of findings. *Evaluation and Program Planning*, *6*, 211–236.

Leff, H. S., & Mulkern, V. (2002). Lessons learned about science and participation from multi-site evaluations. In J. M. Herrell & R. B. Straw (Eds.), *Conducting multiple site evaluations in real-world settings. New Directions for Evaluation, no. 94* (pp. 89–100). San Francisco, CA: Jossey-Bass.

Luborsky, L., McLellan, A. T., Diguer, L., Woody, G., & Seligman, D. A. (1997). The psychotherapist matters: Comparison of outcomes across twenty-two therapists and seven patient samples. *Clinical Psychology: Science & Practice*, *4*, 53–65.

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

Macias, C., Propst, R., Rodican, C., & Boyd, J. (2001). Strategic planning for ICCD clubhouse implementation: Development of the Clubhouse Research and Evaluation Screening Survey (CRESS). *Mental Health Services Research*, *3*, 155–167.

Malysiak, R., Duchnowski, A., Black, M., & Greeson, M. (1996). Establishing wrap around fidelity through participatory evaluation. In *Proceedings of the Ninth Annual Research Conference. A System of Care for Children's Mental Health: Expanding the Research Base*.

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

McGrew, J. H., Pescosolido, B., & Wright, E. (2003). Case managers' perspectives on critical ingredients of assertive community treatment and on its implementation. *Psychiatric Services*, *54*, 370–376.

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.

Mills, S. C., & Ragan, T. J. (2000). A tool for analyzing implementation fidelity of an integrated learning system (ILS). *Educational Technology Research and Development*, *48*, 21–41.

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

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

National MH Advisory Council. (1999). *Bridging science and service: A report by the National Advisory Mental Health Council's Clinical Treatment and Services Research Workgroup*.

Nguyen, T. D., Attkisson, C. C., & Stegner, B. L. (1983). Assessment of patient satisfaction: Development and refinement of a service questionnaire. *Evaluation and Program Planning*, *6*, 299–314.

NH-Dartmouth Psychiatric Research Center. (2002, January). Implementing Evidence-Based Practices Project National Meeting. In T. Singer & P. W. Singer (Eds.), *Implementing Evidence-Based Practices Project Newsletter, 1*.

Orwin, R. G. (2000). Assessing program fidelity in substance abuse health services research. *Addiction*, *95*(Suppl. 3), S309–S327.

Paulson, R. I., Post, R. L., Herinckx, H. A., & Risser, P. (2002). Beyond components: Using fidelity scales to measure and assure choice in program implementation and quality assurance. *Community Mental Health Journal*, *38*, 119–128.

Price, R. H., Friedland, D. S., Choi, J., & Caplan, R. D. (1998). Job-loss and work transitions in a time of global economic change. In X. Arriaga & S. Oskamp (Eds.), *Addressing community problems: Psychological research and interventions* (pp. 195–222). Thousand Oaks, CA: Sage.

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

Raudenbush, S., & Bryk, A. (2001). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Rog, D. J., & Randolph, F. L. (2002). A multi-site evaluation of supported housing: Lessons learned from cross-site collaboration. In J. M. Herrell & R. B. Straw (Eds.), *Conducting multiple site evaluations in real-world settings. New Directions for Evaluation, no. 94* (pp. 61–72). San Francisco, CA: Jossey-Bass.

Salyers, M. P., Bond, G. R., Teague, G. B., Cox, J. F., Smith, M. E., Hicks, M. L., & Koop, J. I. (in press). Is it ACT yet? Real-world examples of evaluating the degree of implementation for assertive community treatment. *Journal of Behavioral Health Services and Research*.

Sechrest, L., West, S. G., Phillips, M. A., Redner, R., & Yeaton, W. (1979). Some neglected problems in evaluation research: Strength and integrity of treatments. In L. Sechrest, S. G. West, M. A. Phillips, R. Redner, & W. Yeaton (Eds.), *Evaluation studies review annual* (Vol. 4, pp. 15–35). Thousand Oaks, CA: Sage.

Snijders, T., & Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage.

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.

Szulanski, G., & Winter, S. (2002). Getting it right the second time. *Harvard Business Review*, *80*, 62–69.

Teague, G. B. (2000). Patient perceptions of care measures. In A. J. Rush, H. A. Pincus, et al. (Eds.), *Handbook of psychiatric measures* (pp. 169–194). Washington, DC: American Psychiatric Association.

Teague, G. B., Bond, G. R., & Drake, R. E. (1998). Program fidelity and 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.

Teague, G. B., Ganju, V., Hornik, J. A., Johnson, J. R., & McKinney, J. (1997). The MHSIP mental health report card: A consumer-oriented approach to monitoring the quality of mental health plans. *Evaluation Review*, *21*(3), 330–341.

Unrau, Y. A. (2001). Using client exit interviews to illuminate outcomes in program logic models: A case example. *Evaluation and Program Planning*, *24*, 353–361.

Vincent, M. L., Paine-Andrews, A., Fisher, J., Devereaux, R. S., et al. (2000). Replication of a community-based multicomponent teen pregnancy prevention model: Realities and challenges. *Family and Community Health*, *23*, 28–45.

Weisman, A., Nuechterlein, K. H., Goldstein, M. J., & Snyder, K. S. (2000). Controllability perceptions and reactions to symptoms of schizophrenia: A within-family comparison of relatives with high and low expressed emotion. *Journal of Abnormal Psychology*, *109*, 167–171.

Weisman, A., Tompson, M. C., Okazaki, S., Gregory, J., et al. (2002). Clinicians' fidelity to a manual-based family treatment as a predictor of the one-year course of bipolar disorder. *Family Process*, *41*, 123–131.

Yates, B. T. (1994). Toward the incorporation of costs, cost-effectiveness analysis, and cost-benefit analysis into clinical research. *Journal of Consulting and Clinical Psychology*, *62*, 729–736.

Yates, B. T. (1995). Cost-effectiveness analysis, cost-benefit analysis, and beyond: Evolving models for the scientist-manager-practitioner. *Clinical Psychology: Science & Practice*, *2*, 385–398.

Yates, B. T. (1996). *Analyzing cost, procedures, processes, and outcomes in human services*. Thousand Oaks, CA: Sage.