

A META-ANALYTIC APPROACH TO GROWTH CURVE ANALYSIS^{1,2}

AURELIO JOSÉ FIGUEREDO

*Evaluation Group for Analysis of Data, Department of Psychology
University of Arizona*

AUDREY J. BROOKS

*Evaluation Group for Analysis of Data, Department of Psychology
University of Arizona*

Southern Arizona Veterans Administration Health Care System

H. STEPHEN LEFF

*Human Services Research Institute, Harvard Medical School
Department of Psychiatry at Cambridge Hospital, Cambridge, Massachusetts*

LEE SECHREST

*Evaluation Group for Analysis of Data, Department of Psychology
University of Arizona*

Summary.—A meta-analytic approach to growth curve analysis is described and illustrated by applying it to the evaluation of the Arizona Pilot Project, an experimental project for financing the treatment of the severely mentally ill. In this approach to longitudinal data analysis, each individual subject for which repeated measures are obtained is initially treated as a separate case study for analysis. This approach has at least two distinct advantages. First, it does not assume a balanced design (equal numbers of repeated observations) across all subjects; to accommodate a variable number of observations for each subject, individual growth curve parameters are differentially weighted by the number of repeated measures on which they are based. Second, it does not assume homogeneity of treatment effects (equal slopes) across all subjects. Individual differences in growth curve parameters representing potentially unequal de-

¹This research was supported in part by the Arizona Division of Behavioral Health Services (ADBHS), by a grant from the National Institute of Mental Health (Grant #5 R01 MH44878-02), and by a grant from the National Institute on Drug Abuse (Grant #1 R18 DA06918-01). The authors thank the staffs of the capitated and fee-for-service programs, and Kendis Stake, Boyd Dover, and Ted Williams of the ADBHS for providing the cooperation and support this research required. We also thank members of the Outside Assessment Teams and colleagues at the Human Services Research Institute for their work in collecting the data. We thank Lewis Petrinovich, David Rogosa, and Keith Widaman for much informative discussion and guidance on both the theoretical rationale underlying and the practical techniques for performing growth curve analysis, and Bradley Smith for helpful comments on the manuscript. We also appreciate the help of Mike Berren and Carmen Valiente, who were instrumental in developing the system of orthogonal contrasts used to operationalize *a priori* hypotheses regarding the Pretreatment Resource Allocation Functional Level Scale scores.

²Please address correspondence to Aurelio José Figueredo, Department of Psychology, University of Arizona, Tucson, AZ 84721-0068.

velopmental rates through time are explicitly modeled. A meta-analytic approach to growth curve analysis may be the optimal analytical strategy for longitudinal studies where either (1) a balanced design is not feasible or (2) an assumption of homogeneity of treatment effects across all individuals is theoretically indefensible. In our evaluation of the Arizona Pilot Project, individual growth curve parameters were obtained for each of the 13 rationally derived subscales of the New York Functional Assessment Survey, over time, by linear regression analysis. The slopes, intercepts, and residuals obtained for each individual were then subjected to meta-analytic causal modeling. Using factor analytic models and then general linear models for the latent constructs, the growth curve parameters of all individuals were systematically related to each other via common factors and predicted based on hypothesized exogenous causal factors. The same two highly correlated common factors were found for all three growth curve parameters analyzed, a general psychological factor and a general functional factor. The factor patterns were found to be nearly identical across the separate analyses of individual intercepts, slopes, and residuals. Direct effects on the unique factors of each subscale of the New York Functional Assessment Survey were tested for each growth curve parameter by including the common factors as hierarchically prior predictors in the structural model for each of the indicator variables, thus statistically controlling for any indirect effect produced on the indicator through the common factors. The exogenous predictors modeled were theoretically specified orthogonal contrasts for Method of Payment (comparing Arizona Pilot Project treatment or "capitation" to traditional or "fee-for-service" care as a control), Treatment Administration Site (comparing various locations within treatment or control groups), Pretreatment Assessment (comparing general functional level at intake as assigned by an Outside Assessment Team), and various interactions among these main effects. The intercepts, representing the initial status of individual subjects on both the two common factors and the 13 unique factors of the subscales of the New York Functional Assessment Survey, were found to vary significantly across many of the various different treatment conditions, treatment administration sites, and pretreatment functional levels. This indicated a severe threat to the validity of the originally intended design of the Arizona Pilot Project as a randomized experiment. When the systematic variations were statistically controlled by including intercepts as hierarchically prior predictors in the structural models for slopes, recasting the experiment as a nonequivalent groups design, the effects of the intercepts on the slopes were found to be both statistically significant and substantial in magnitude. Furthermore, the contrasts for Pretreatment Assessment scores also predicted statistically significant proportions of variance in both the two common factors and the 13 unique factors of the subscales of the New York Functional Assessment Survey for all three growth curve parameters, confirming an influence of the initial status of individual subjects on treatment effect. This empirical example illustrates both the mechanics and the many practical benefits of a meta-analytic approach to growth curve analysis in program evaluation.

The analysis of data from longitudinal studies often involves making various difficult choices. These choices may involve complex and sometimes conflicting considerations related to whether measures are obtained at the same times for every individual, the extent of any autoregressive effects, whether correlations between adjacent times are equal (stationary) for all pairs, and whether data are missing for any individuals at any measurement

occasions. Depending on the resolution of these issues, various possible approaches to the analysis of change data, such as repeated-measures analysis of variance, random effects regression models, and panel design path analysis, may or may not be appropriate. An analytic approach that has gained popularity in recent years is Growth Curve Analysis. Paradoxically, due to the explosive growth of this field, the longitudinal researcher now faces an embarrassment of riches when it comes to selecting from among the various quantitative models that are now available. These alternative models include such sophisticated methodologies as the longitudinal applications of Hierarchical Linear Models (e.g., Bryk & Raudenbush, 1992) and Structural Equation Models (e.g., Meredith, 1991).

The purpose of this paper is to present yet another alternative, which we refer to as the meta-analytic approach to growth curve analysis, and to describe both its advantages and disadvantages in relation to the more frequently used methods. This paper is not intended as either a comprehensive review of the state of the art in this area or a further elaboration of the common mathematical theory underlying all of these models, but instead as a contribution to the armamentarium of analytic tools available to the practicing data analyst. We make no claim that our approach is better in some global sense than its alternatives but merely suggest that under certain specifiable conditions its advantages may outweigh its limitations in relation to the other possibilities. Thus, we hope to make the case that under those particular circumstances, a meta-analytic approach to growth curve analysis might actually become the method of choice in spite of its relative simplicity in relation to the more sophisticated methods now available. We start with a brief summary and description of the general advantages of growth curve analysis over more traditional methods.

Growth curve analysis refers to a way of representing data in terms of change over time and consists simply of two-dimensional plots of the value of a variable of interest against time. A more generic term would be "change curves," since one need not assume any particular form for the plots that result, and the term "growth" carries with it implications of some sort of organic, inherent process. In the context of change by persons with severe mental illness receiving rehabilitation services, "rehabilitation curves" would be a more precise description (Leff, 1992). However, in this paper we continue to use the term "growth curves" in the interests of continuity. Growth curves may be used to represent the data for any unit of observation or analysis from the individual on up through subgroups to a population. The familiar curves meant to indicate average growth rates for children, for example, are representations of population data.

An important advantage of growth curves is that they can be plotted for individuals and, thus, preserve the observations at the individual level. As is

now well known, average growth curves do not necessarily estimate well the growth of any single individual; they are smoothed by the averaging out of many individual deviations. A second advantage of growth curves is that they do not necessarily require measures of individuals at the same times nor even of the same number of time points per individual (although some of the more sophisticated analytic models do indeed require these restrictions). That is to say, individual growth curves are, at least in theory, relatively robust against problems involving missing data. A third advantage is that growth curves may be more sensitive to emerging differences (trends) between new programs and established controls than simple *pre* and *post* comparisons when the follow-up period is relatively short (Leff, Mulkern, Lieberman, & Raab, 1996).

Any given growth curve may be described by a large number of different parameters, although in practice most information of interest is likely to be captured by only a few such parameters. Specifically, those parameters likely to be most useful would be intercept, slope, and dispersion. The intercept estimates what the value of the dependent variable was at the time of the first observation. One could, then, compare growth curves for two units, individuals or groups and determine whether they differed at the beginning of observations or, by extrapolation, at some earlier time. The slope estimates the growth rate or rate of change, the amount of change in the dependent variable that is occurring with each successive unit of time. The slope may be positive, zero, or negative, depending on whether the variable of interest is increasing, unchanging, or decreasing with time. The magnitude of the slope will indicate whether the change is large or small. Dispersion is a characteristic of the data rather than the growth curve itself, but the parameter indicates how well the curve fits the data by describing the variability around the line defining the slope.

We would like to emphasize that growth curve analysis is not some sort of exotic statistical *legerdemain*. Analysis of growth curves is reducible to well-understood statistical models in the general family of linear regression models (Laird & Ware, 1982; Gibbons, Hedeker, Waternaux, & Davis, 1988; McCauley & Anderson, 1989; Bryk & Raudenbush, 1992). Warnings are often issued against the hazards of correlating observations with time because of correlated error terms (autoregression), but McCain and McCleary (1979) have pointed out that the problems associated with autoregression have to do mostly with calculating the correct error terms for the purpose of separately assessing the statistical significance of the growth curve parameters of individual subjects. McCain and McCleary go on to note that the estimates of the growth curve parameters obtained by correlating variables with time will be unbiased in spite of this apparent limitation. The meta-analytic approach to growth curve analysis being described here does not con-

trol for serially autoregressive effects. However, our proposed approach does not require the calculation of statistical significance for the growth curve parameters of different individuals and is, therefore, largely unaffected by this limitation.

The parameters just listed may be calculated for every growth curve obtained from a study or data set. Thus, if a study has 80 cases, one could have 80 growth curves and 80 sets of parameter estimates. What that means is that the data for each individual case are reduced to a small number of descriptors of whatever change occurs over time. The alternative would be to combine data across individuals at each time and estimate an overall growth curve. That has the important disadvantage of obliterating information at the individual level such as differences in rates of change associated with individual characteristics.

Calculating growth curve parameters for individual cases in effect transforms each case into an individual study. Thus, if two subjects are exposed to an intervention meant to increase self-esteem over time, treating their data as individual cases means that each of them constitutes a test of the effectiveness of self-esteem therapy. That conception renders the modeling of growth curve parameters equivalent in nature to meta-analysis. Thus, one can combine the information from a whole set of growth curves to achieve a more coherent and powerful analysis than would be permitted by any one or small set of the individual analyses, just as one wants to combine information from separate experiments.

This meta-analytic conception of growth curve analysis brings into play a wide range of strategies for data analysis in which growth curve parameters are treated as dependent variables, just as effect size estimates are treated as dependent variables in meta-analysis. One can determine whether systematic relationships exist between such characteristics of individuals as group membership, sex, diagnostic status, and so on, are related to intercept, slope, residual, and other growth curve parameters. One can also enter those latter parameters into hierarchical analyses to determine, for instance, whether after removing variance attributable to the intercept parameter (initial difference), any variation in slope (rate of change) remains. We believe that this meta-analytic approach to growth curve analysis provides a conceptually simple and analytically powerful tool for dissecting the complexities of change over time. The techniques used in this paper are primarily based on the individual regression approach proposed by Petrinovich and Widaman (1984) for the analysis of repeated-measures data but were also greatly influenced by the ideas and perspectives of Rogosa (personal communication, 1991; Rogosa & Willett, 1985).

To illustrate both the details of these analytic procedures and the kind of results obtainable from fairly typical data, we report here a meta-analytic

strategy that was successfully applied to growth curve data from an evaluation of the Arizona Pilot Project, an experimental project for financing the treatment of the severely mentally ill. The major substantive results from that project are reported elsewhere (Leff, Mulkern, Lieberman, & Raab, 1994; Leff, *et al.*, 1996). These results were qualitatively similar to those of our growth curve analysis, but the methodological differences and their implications for the interpretation of the basic findings will be highlighted as each separate issue is discussed. Furthermore, the past few years have seen the publication of a series of excellent papers on the application of growth curve analysis and related methods to tracking the rehabilitation of the seriously and persistently mentally ill (e.g., Ryan, Sherman, & Judd, 1994; Barkham, Rees, Shapiro, Hardy, Stiles, & Reynolds, 1997; Brekke, Long, Nesbitt, & Sobel, 1997; Ryan, Sherman, & Bogart, 1997). In this paper, we focus primarily on the methodological issues associated with the meta-analytic approach to growth curve analysis.

METHOD

Design of the Experiment

The Arizona Pilot Project was planned as a randomized experiment to test the effects on severely mentally ill patients of capitation as a method of reimbursing their service providers, with the usual method of fee-for-service care as a comparison. Mentally ill persons within treatment facilities were to be assigned randomly to capitation or usual reimbursement groups. Under capitation, the treatment unit was paid a flat amount of money to provide services to each patient for one year, an important *proviso* being that each patient would be assigned to a case manager charged with seeing to it that the patient received necessary services. The treatment unit stood to lose money if costs during the year exceeded capitation revenues but could use savings for other purposes if costs could be kept below revenues. The treatment units negotiated for any services required from other agencies and paid for them directly. Under the usual reimbursement scheme, patients were determined by treatment units to need services of one sort or another, which had then to be arranged with an appropriate agency, and that agency billed the state for any services provided.

All patients entered into the study were assessed at baseline for functional level and then quarterly afterwards for a period of up to six quarters. A variety of different measures were employed, but we will focus in this paper on the New York Functional Assessment Survey, which was filled out quarterly on each patient by the patient's case manager. Some cases had missing data in one or more quarters in the middle of the sequence because case managers failed to complete forms. Some patients were lost to follow-up with data missing at the end of the follow-up period.

All patients who met the Arizona definition of severely mentally ill in the study sites were placed in blocks defined by functional level, diagnosis, and age. Patients were then randomly assigned to conditions from blocks according to a specified sampling plan. If a patient refused enrollment in a capitated program, another patient from the same block was enrolled. However, some deviation from the sampling plan occurred if a block became empty (Leff, Mulkern, Raab, & White, 1991). Although assignment was meant to have been random, from strata defined by functional level, diagnosis, and age, it appears that randomization was not completely maintained. We conclude this in part from the analyses that follow but also from the fact that patients were permitted to refuse enrollment in the capitated program (Leff, *et al.*, 1991). Nonetheless, failures of randomization were not massive, and the number of cases was, as such experiments go, large. The study was carried out in the two major counties in Arizona, Maricopa County and Pima County, with a total of six different mental health agencies being involved.

Functional Assessment Measures

The New York Functional Assessment Survey (Furman & Lund, 1979) is an instrument intended to measure the level of functioning of mental patients consisting of 81 individual descriptive items on which patients may be rated by case managers. The individual items described psychological symptoms (e.g., anxiety, hostile behavior) as well as daily functioning (e.g., brush teeth, do laundry). The items were rated on a Likert scale ranging from 1 (no problem) to 4 (severe problem) for the psychological symptoms and 1 (can not or will not) to 3 (adequately and consistently) for the daily functioning variables. However, some New York Functional Assessment Survey items were recoded so that higher scores would always indicate more severe problems.

The items from the New York Functional Assessment Survey were theoretically assigned to 13 separate subscales based on item content. The verbal labels for these subscales were as follows: (1) Potentially Harmful Behavior, (2) Antisocial Behavior, (3) Anxiety and Depression, (4) Thinking Disorder, (5) Uncooperativeness and Suspiciousness, (6) Withdrawal and Retardation, (7) Community Living Skills, (8) Personal Care Skills, (9) Nutrition Habits, (10) Home Care Skills, (11) Interpersonal Skills, (12) Community Recreation Use, and (13) Health Care Skills. These subscales then became the raw material for our subsequent analyses, modeling them with two higher-order common factors, one related to the psychological symptoms of mental disorder and the other to global level of functioning.

In addition to the New York Functional Assessment Survey, all patients were assessed at baseline by an Outside Assessment Team, *i.e.*, a team not otherwise involved in the study, on the Resource Allocation Functional Level

Scale (Leff, *et al.*, 1996). This assessment was for the purpose of establishing baseline status and was based on careful examination of the case record for each patient. The scoring criteria for this seven-point scale were as follows: (1) Dangerous, (2) Unable to Function Due to Current (Acute) Psychiatric Symptoms, (3) Lacks Personal Care Skills, (4) Lacks Community Living Skills, (5) Needs Role Support and Training, (6) Needs Support or Treatment to Cope with Extreme Stress or to Maintain or Enhance Personal Development, and (7) System Independent. The Outside Assessment Team members also completed the New York Functional Assessment Survey for a sample of patients at baseline on the basis of case record information. Comparisons of New York Functional Assessment Survey ratings by the Outside Assessment Team and Case Managers indicated acceptable interrater reliability (Leff, *et al.*, 1991).

Statistical Analysis

Statistical analyses were performed using various univariate procedures in the Statistical Analysis System (SAS Institute, Inc., 1989). Individual regressions were performed using Regression Procedure, exploratory factor analyses using Factor Procedure, and hierarchical general linear modeling of sample growth curve parameters using General Linear Models Procedure.

Level 1 growth curve analysis.—Individual growth curve parameters were obtained by bivariate linear regression for each of the variables of interest over time. Initially, each subject on which repeated measures were obtained was implicitly treated as a separate case study for analysis. Thus, the following regression parameters were estimated for each individual subject: (a) Intercept, representing our best estimate of the initial condition of each subject, (b) Slope, or unstandardized regression weight, representing the direction and magnitude of average change in status for each individual subject over time (as manifested in quarterly assessments by case managers), and (c) Residual, or root-mean-squared error, representing the intraindividual variability of each subject about that individual trend or individual instability of temporal progress. We made no attempt to estimate nonlinear parameters for the data, although this approach can be readily applied to curvilinear regression. Thus, individual regression analyses were performed for each subject on each of the 13 rationally-derived subscales of the New York Functional Assessment Survey. This initial step is called the Level 1 growth curve analysis. These parameter estimates, or chronometric constructs, were then used as data for further statistical modeling, which is called the Level 2 growth curve analysis (Willett & Sayer, 1994). This two-step procedure can be used to estimate with great precision all the principal parameters of a hierarchical linear model (Rogosa, 1989; Bryk & Raudenbush, 1992; Rogosa & Saner, 1994).

This approach had at least three distinct advantages over its more traditional alternatives, Repeated Measures Analysis of Variance and "Panel Design" Path Analysis. First, it did not assume a balanced sampling design (equal numbers of repeated observations) across all subjects. Most sophisticated chronometric models currently available for the application of structural equation models to growth curve analysis still appear to require balanced sampling design and equal intervals between repeated observations. Due to attrition over time, however, it was unreasonable to expect that equal numbers of observations would be obtained for each individual. Furthermore, it would have been wasteful to discard the data for any individuals for whom information was not complete but dubious to forecast values for them at any times that they were not observed. Instead, individual growth curve parameters were differentially weighted by the number of repeated measures they were based on, as is typically done in meta-analytic models of multiple studies (Hedges & Olkin, 1985).

The second major advantage was that this analysis, as a random effects model (cf. Laird & Ware, 1982; Gibbons, *et al.*, 1988; Bryk & Raudenbush, 1992), did not assume homogeneity of rates or change (equal slopes) across all subjects because it was also deemed unreasonable to assume equal recovery rates over time across all individuals. Instead, individual differences in growth curve parameters, representing potentially unequal rates of change through time, were explicitly modeled (cf. Rogosa, Brandt, & Zimowski, 1982). Thus, growth curve analysis can often be the optimal analytical strategy for longitudinal studies, such as this one, in which an assumption of homogeneity of recovery rates across all individuals is theoretically implausible.

The third major advantage of this method was that growth curve parameters, based on all the observations available for each individual, are generally much more reliable than the scores measured at each cross-sectional measurement occasion. This is true for intercepts as well as slopes. Thus, although such data aggregation superficially reduces the available degrees of freedom for analysis, it has been shown actually to enhance the overall statistical power due to the substantial reduction in errors of measurement (Sutcliffe, 1980; Petrinovich & Widaman, 1985). Furthermore, this enhanced reliability can be expected to have the added benefit of disattenuating any correlations between variables (Pedhazur, 1982; Cohen & Cohen, 1983).

Level 2 growth curve analysis.—The Level 2 growth curve analysis consisted of two successive components: the measurement model and the structural model. The measurement model is the portion of the model wherein a number of directly observed measures are related to a smaller set of hypothetical constructs that cannot be directly observed but are nonetheless the latent entities of ultimate theoretical interest. The structural component of

the model is the causal analysis of the hypothetical constructs that were produced by the measurement models. Thus, the Level 2 chronometric constructs represented by both the initial status factors and the change factors were systematically related to hypothesized predictors as specified by a *a priori* causal theory.

The measurement model.—Using exploratory factor analytic procedures, the chronometric constructs representing the growth curve parameters of all individuals were systematically related to each other via higher-order common factors, hierarchically modeling the covariances among chronometric parameters with higher-order psychometric constructs or Level 2 chronometric constructs. Exploratory factor analyses of the entire sample were performed separately for each of the three individual regression parameters described above. Thus, separate common factors were constructed for the intercepts, the slopes, and the residuals of the 13 New York Functional Assessment Survey subscales. Common factor modeling was performed using unstandardized covariance matrices, squared multiple correlations as prior commonality estimates, subjective scree test for retaining a reduced number of factors, and oblique (Promax) factor rotation.

This procedure was preferable to the successive factor analysis of each temporal cross-section of the data because, at least in principle, it would have been able to distinguish the change factors underlying the covariances between the slopes of different variables from the initial status factors underlying the covariances between their intercepts (for a detailed discussion of these issues, see the edited monograph by Collins & Horn, 1991). This distinction is said to be important because the causal processes that shaped the multivariate patterns defining the initial status of individuals might not be the same as those that operate afterwards (Cunningham, 1991; Meredith, 1991). Panel designs based on factor models of successive temporal, cross-sections monitor quantitative changes in the successive values of the common factors. These common factors, however, represent fixed multivariate patterns that are typically constrained to be qualitatively invariant over time (Cunningham, 1991; Widaman, 1991). Factor models of growth curve parameters instead may capture the dynamic multivariate pattern of the change itself, rather than the change in some fixed multivariate pattern, and thus create a multivariate profile of the different causal processes that may operate throughout treatment (McArdle, 1988; McArdle & Hamagami, 1991; Nesselroade, 1991). By constructing separate common factors for the intercepts, slopes, and residuals it was also possible to determine what predictive relationship the multivariate patterns of initial status had with those of subsequent treatment. For example, the Intercept Factors represented the static patterns measured at a single point in time, but the Slope Factors indicated any changes that might have occurred with treatment in the composition of

symptomatology over time. The latent structure of such transformations can best be represented by change factors. From a clinical point of view, it may be that using change factors, rather than fixed multivariate patterns, better represents concepts like course and prognosis (Leff, 1992).

Although certain more sophisticated methods, such as Hierarchical Linear Models (Bryk & Raudenbush, 1992), are currently available for the modeling of growth curves, these suffer from the disadvantage of ultimately remaining at the level of a univariate analysis in that they lack a multivariate measurement model. Whereas it is always possible to aggregate factor scores within cross-sectional subsamples before estimating the growth curves of the composited factor scores themselves with a Hierarchical Linear Model, our two-step meta-analytic approach is more suitable for addressing the issue of whether the multiple indicators are indeed changing together over time. In the context of present data, it is difficult to make a case for the importance of this advantage because our convergent results paradoxically support the validity of the same factors over time. Nevertheless, it is arguably a methodological advantage to be able to determine whether the common factors actually represent covariation over time and not just within successive cross-sectional slices of the data.

The structural model.—Using general linear models, the constructs represented by both the initial status factors and the change factors were systematically related to hypothesized predictors as specified by *a priori* causal theory. This approach avoided the estimation, otherwise necessary (Cook & Campbell, 1979), of all the serially autoregressive effects of successive measures of the same variable on each other, and concentrated our statistical power on how the changes over time represented by one latent variable influenced the changes represented by another.

As noted above, this procedure also permitted the testing of hypotheses regarding how initial status constructs might influence the chronometric constructs, or change factors, both within the same variable and between different ones. Thus, although the proposed study was started with some of the formal characteristics of a Randomized Clinical Trial, such as random assignment of subjects to treatment groups, sampling was destined to almost immediately become quasi-random due to differential recruitment and attrition within different treatment programs (cf. Cook & Campbell, 1979; Leff, *et al.*, 1991, 1994, 1996; Shadish, Cook, & Leviton, 1991; Sechrest & Figueredo, 1993). To counter this threat, we operationalized this analysis from the outset as a quasi-experimental design with nonequivalent groups (Cook & Campbell, 1979). The effects of both the treatments and other hypothesized predictors on the outcome factors were statistically controlled for the effects of possible systematic differences in their baseline levels for the different

groups. These precautions would not have been necessary in a true Randomized Clinical Trial design.

General linear modeling was performed separately for each of the three individual regression parameters described above on both the common factors and the unique residual factors of the 13 New York Functional Assessment Survey subscales. The residual effects on the unique factors were obtained by including the relevant common factor score as a hierarchically prior predictor in the general linear model for each of the indicator variables, thus statistically controlling for any indirect effect produced through the relevant common factor.

Other predictors modeled were Capitation (a single contrast comparing Arizona Pilot Project method of payment to fee-for-service care as a control), Treatment Administration Site (four orthogonal contrasts comparing various treatment locations both across and within method of payment groups), and Pretreatment Resource Allocation Functional Level Scale (four orthogonal contrasts comparing ratings on a seven-point ordinal scale of general functional level as assigned by an Outside Assessment Team at intake), and various interactions among these main effects. Thus, the following system of orthogonal contrasts was constructed. The *Capitation* contrast compared all patients receiving services under capitation with those receiving fee-for-service care. The *County* contrast compared all patients in Maricopa County with those in Pima County. The *Capitation vs County* contrast represents the interaction of the two previous contrasts. The *Pima Capitation Sites* contrast compared patients receiving services under capitation at the two treatment administration sites within Pima County. The *Maricopa Capitation Sites* contrast compared patients receiving services under capitation at the two treatment administration sites within Maricopa County. The *Community Skills* contrast compared all patients having community skills with those lacking community skills. The *Functional Skills* contrast compared patients having functional skills but lacking community skills with those lacking both functional and community skills. The *Personal Skills* contrast compared patients having both functional and personal skills but lacking community skills with those having functional skills but lacking both personal and community skills. The *Role Support* contrast compared patients having both community skills and role support with those having community skills but lacking role support.

The general linear models described above were used to control statistically for the significant effects of differences in individual intercepts on individual slopes, and for the significant effects of differences in both individual intercepts and slopes on individual root mean squared errors. This was done by (a) including the intercept of each common factor as a hierarchically prior predictor in the general linear model for the slope of that

common factor, (b) including both the intercept and slope of each common factor as hierarchically prior predictors in the general linear model for the root mean squared error of that common factor, (c) including the intercept of each indicator, as well as that of the relevant common factor, as hierarchically prior predictors in the general linear model for the slope of that indicator, and (d) including both the intercept and slope of each indicator, as well as those of the relevant common factor, as hierarchically prior predictors in the general linear model for the root mean squared errors of that indicator.

The hierarchical general linear models contained all the predictors and interactions that were hypothesized *a priori* in their prespecified order of causal priority. These inclusive models were used for the initial significance testing of statistical hypotheses. However, to maximize the efficiency of parameter estimation (Cohen & Cohen, 1983), all the models were afterwards respecified to include only those predictors found statistically significant ($p < .05$) within the prior hierarchical analyses. The only exceptions to this general practice were the inclusion of any nonsignificant main effects where interaction terms containing them were found significant and of lower-order interactions where higher-order interaction terms containing them were found significant. Thus, the restricted regression models contained only the truly relevant variables: all parameter estimates reported were obtained by the simultaneous ordinary least-squares estimation of the restricted models. Because the statistical significance of predictors was determined in advance by the theoretically specified models, these later model respecifications enhanced the final efficiency of parameter estimation without incurring any risk of capitalization on chance.

RESULTS

The same two highly correlated common factors were found for all three regression parameters analyzed: a general psychological factor and a general functional factor. Both the factor patterns and the factor intercorrelations, displayed in Table 1, were found to be nearly identical across the separate analyses of individual intercepts, slopes, and errors. This implies, for example, that the hypothesized Change Factors were not in fact different from the Initial Status Factors. These results do not demonstrate that the theoretical arguments outlined above for the conceptual distinction between such constructs are generally wrong but merely that they happen to have no bearing on these particular data. In Table 1, the factor loadings are expressed as standardized regression coefficients (β -weights), and the factor intercorrelations as bivariate Pearson product-moment correlation (r) coefficients.

The Intercept Factors, representing the initial functional status of indi-

TABLE 1
 FACTOR PATTERNS (β -WEIGHTS) AND FACTOR INTERCORRELATIONS
 FOR INTERCEPT, SLOPE, AND RESIDUAL FACTORS

Subscale*	Intercept Factors		Slope Factors		Residual Factors	
	Functional	Psycho-logical	Functional	Psycho-logical	Functional	Psycho-logical
1	-.09	.72	.01	.48	.05	.47
2	.28	.53	.01	.46	.22	.36
3	-.11	.56	-.00	.50	-.06	.47
4	.42	.42	-.03	.68	.14	.53
5	.11	.69	-.02	.66	.04	.62
6	.42	.33	.07	.53	.06	.47
7	.82†	.02	.57	.08	.36	.12
8	.83	-.10	.56	.06	.74	-.04
9	.37	.02	.34	.03	.37	.12
10	.89	-.12	.65	-.11	.76	-.09
11	.69	.21	.55	.11	.57	.14
12	.74	.04	.56	-.00	.48	.09
13	.82	.02	.59	.02	.60	.03
Functional	1.00		1.00		1.00	
Psychological	.60	1.00	.50	1.00	.57	1.00

*1: Potentially Harmful Behavior, 2: Antisocial Behavior, 3: Anxiety and Depression, 4: Thinking Disorder, 5: Uncooperativeness and Suspiciousness, 6: Withdrawal and Retardation, 7: Community Living Skills, 8: Personal Care Skills, 9: Nutrition Habits, 10: Home Care Skills, 11: Interpersonal Skills, 12: Community Recreation Use, 13: Health Care Skills.

†Loadings >.30.

vidual subjects, were found to vary significantly across many of the various different method of payment conditions, treatment sites, and pretreatment functional levels. This was true for both common factors of the New York Functional Assessment Survey. These intercept differences indicated a severe threat to the validity of the originally intended design of the evaluation as a randomized experiment. The general linear models predicting the intercept factors are shown in Table 2. The hierarchical and overall tests of significance, i.e., the F ratios, degrees of freedom, and associated probabilities under the null hypothesis, were obtained from the inclusive hierarchical analyses; the parameter estimates, expressed as standardized regression coefficients, and the squared multiple correlations were obtained from the restricted models by simultaneous ordinary least-squares estimation using only the predictors found statistically significant by the prior hierarchical analyses. Exact probabilities were reported following Rosenthal (1969), who described the informative uses of exact probabilities in meta-analysis and the extra affordances that they provide to future researchers who might be applying varying or changing conventions of statistical significance. Nevertheless, the decision rule that we applied was based upon the current conventional alpha level of $p < .05$. As described above, nonsignificant main effects were retain-

TABLE 2
FINAL REGRESSION MODELS FOR INTERCEPT FACTORS OF
PSYCHOLOGICAL AND FUNCTIONAL SYMPTOMS

Intercept	β -weight	$F_{1,826}$	$P(H_0)$
Psychological			
Capitation	-.09	0.15	.7024
County	.02	0.08	.7741
Capitation \times County	.06	4.15	.0420
Pima Capitation Sites	-.05	1.19	.2764
Maricopa Capitation Sites	-.15	47.03	.0001
Community Skills	-.46	122.53	.0001
Functional Skills	-.24	65.00	.0001
Personal Skills	-.06	7.91	.0050
Role Support	-.17	30.04	.0001
Capitation \times Community Skills	.07	7.72	.0056
Capitation \times Personal Skills	-.09	9.18	.0025
Pima Capitation Sites \times Functional Skills	.09	9.19	.0025
Squared Multiple Correlation	.27	10.85	.0001
Functional		$F_{1,825}$	
Psychological Intercept	.58	960.10	.0001
Capitation	.04	3.27	.0709
County	.07	7.30	.0070
Capitation \times County	-.02	1.33	.2486
Maricopa Capitation Sites	-.10	26.11	.0001
Community Skills	-.25	115.33	.0001
Functional Skills	.10	4.98	.0260
Personal Skills	-.13	31.74	.0001
Capitation \times Functional Skills	-.06	9.88	.0017
County \times Functional Skills	-.01	1.31	.2525
Capitation \times County \times Functional Skills	-.07	4.58	.0326
Maricopa Capitation Sites \times Personal Skills	.07	10.88	.0010
Squared Multiple Correlation	.58	39.90	.0001

ed when their interactions were significant; nonsignificant lower-order interactions were retained when their higher-order interactions were significant.

The slope and residual factors, representing the effects of treatment and the unsystematic intraindividual variation of individual subjects across time, were also found to vary significantly across many of the various different method of payment conditions, treatment administration sites, and pretreatment functional levels. When the systematic variations in intercept described above were statistically controlled for in the general linear models predicting the slopes and residuals, thus recasting the experiment as a nonequivalent groups design (Cook & Campbell, 1979), their effects were found to be substantial, and the pattern of results was affected dramatically. This strong effect was found for both common factors of the New York Functional Assessment Survey. The general linear models predicting the slope and residual

factors are shown in Tables 3 and 4, respectively. As in Table 2, the hierarchical and overall tests of significance, i.e., the F ratios, the degrees of freedom, and associated probabilities under the null hypothesis, were obtained from the inclusive models; the parameter estimates, expressed as standardized regression coefficients (β -weights), and the squared multiple correlations were obtained from the restricted models by simultaneous ordinary least-squares estimation using only the predictors found statistically significant by the prior hierarchical analyses, as described above.

TABLE 3

FINAL REGRESSION MODELS FOR SLOPE FACTORS OF PSYCHOLOGICAL AND FUNCTIONAL SYMPTOMS

	Slope	β -weight	$F_{1,824}$	$P(H_0)$
Psychological				
Psychological Intercept		-.58	380.91	.0001
Capitation		-.17	2.66	.1033
County		.04	2.89	.0897
Capitation \times County		.08	3.29	.0702
Maricopa Capitation Sites		.04	1.11	.2921
Community Skills		-.14	12.98	.0003
Functional Skills		-.05	0.18	.6745
Personal Skills		.06	4.05	.0446
Capitation \times Community Skills		.13	5.11	.0240
Capitation \times Functional Skills		.14	9.30	.0024
County \times Functional Skills		.01	0.83	.3613
Capitation \times County \times Functional Skills		-.08	3.77	.0526
Maricopa Capitation Sites \times Community Skills		-.07	4.86	.0277
Squared Multiple Correlation		.34	14.51	.0001
Functional				
			$F_{1,823}$	
Psychological Intercept		.33	137.02	.0001
Functional Intercept		-.47	103.45	.0001
Psychological Slope		.67	539.90	.0001
County		.02	0.00	.9445
Maricopa Capitation Sites		-.03	1.33	.2484
Community Skills		-.16	29.53	.0001
Functional Skills		.01	0.02	.8985
Personal Skills		-.10	15.57	.0001
County \times Community Skills		-.06	5.68	.0174
Maricopa Capitation Sites \times Community Skills		.06	8.63	.0034
Maricopa Capitation Sites \times Functional Skills		-.05	3.92	.0482
Squared Multiple Correlation		.50	26.72	.0001

The theoretically specified orthogonal contrasts for baseline assessment of functional level (Pretreatment Resource Allocation Functional Level Scale scores) were found to predict differentially significant proportions of variance in the intercepts, slopes, and residuals of both common factors of the New York Functional Assessment Survey. There were also substantial varia-

TABLE 4

FINAL REGRESSION MODELS FOR RESIDUAL FACTORS OF PSYCHOLOGICAL AND FUNCTIONAL SYMPTOMS

Residual	β -weight	$F_{1,760}$	$P(H_0)$
Psychological			
Psychological Intercept	.58	322.45	.0001
Functional Intercept	.11	9.15	.0026
Psychological Slope	.25	54.20	.0001
Capitation	.21	60.40	.0001
County	-.20	32.54	.0001
Capitation \times County	-.06	8.26	.0042
Pima Capitation Sites	-.11	18.02	.0001
Maricopa Capitation Sites	.11	18.42	.0001
Community Skills	-.06	0.03	.8649
Functional Skills	-.04	0.69	.4048
Capitation \times Functional Skills	.06	0.06	.8078
County \times Community Skills	.09	7.24	.0073
County \times Functional Skills	.01	1.41	.2349
Capitation \times County \times Functional Skills	-.10	8.69	.0033
Squared Multiple Correlation	.41	16.99	.0001
Functional			
		$F_{1,759}$	
Psychological Intercept	-.28	283.19	.0001
Functional Intercept	.43	190.88	.0001
Psychological Slope	-.14	30.98	.0001
Functional Slope	.16	19.54	.0001
Psychological Residual	.67	638.86	.0001
Capitation	-.00	0.19	.6636
Pima Capitation Sites	-.07	9.55	.0021
Maricopa Capitation Sites	-.02	2.56	.1098
Community Skills	.04	0.11	.7347
Functional Skills	.00	2.02	.1552
Personal Skills	.05	1.67	.1964
Capitation \times Functional Skills	.07	6.67	.0100
Capitation \times Personal Skills	-.06	4.21	.0405
Maricopa Capitation Sites \times Community Skills	-.06	7.52	.0063
Squared Multiple Correlation	.58	35.55	.0001

tions between treatment administration sites and interactions between treatment administration sites and pretreatment functional level scores in the intercepts, slopes, and residuals of both common factors of the New York Functional Assessment Survey. These variations indicated that the specific features of treatment administered at different treatment administration sites differentially affected patient subpopulations. Furthermore, because these treatment administration site effects on slope factors were controlled statistically for the intercept factors, these differences were not wholly attributable to differences in the distributions of patients' initial status across different treatment administration sites.

DISCUSSION

This discussion will focus on some of the findings to illustrate methodological considerations in evaluating the efficacy of an intervention. The order of the discussion will correspond to that of the causal priority assigned in the general linear models, sequentially examining the effects upon Intercepts, Slopes, and Residuals.

Differences in Intercepts

Prior to any discussion of the effectiveness of capitation, one must first examine whether differences in symptoms and functioning were present at the outset in the patients in the capitation *versus* fee-for-service conditions, at the different treatment administration sites, and in the two counties. Some researchers have employed the first observation values as a covariate to control for initial status. However, the intercept is a more reliable measure of initial differences in that it is based on all of the individuals' data (the entire regression line) rather than a single measurement (first observation values). Therefore, by using the intercept as a covariate of treatment effectiveness one obtains a more accurate picture of change over time.

It should also be emphasized here that some of the currently predominant approaches to growth curve analysis use the individual subject means, rather than the true intercepts, as a way of controlling for systematic individual differences in absolute levels on the study variables. These predominant approaches include Hierarchical Linear Models and other Random Effects Regression Models as well as most current applications of Structural Equations Models. In the meta-analytic approach to growth curve analysis, we prefer to use actual intercepts because the means are aggregated over time and are therefore contaminated with information from potentially different individual slopes. Thus, if two subjects have the same intercepts (indicating equal baselines) but different slopes (indicating unequal rates of recovery), their means aggregated over time on the criterion variable will be unequal. Individual subject means therefore cannot be used to directly model systematic differences in baseline status because they are necessarily affected by the course of subsequent events over time. Intercepts represent a more accurate measure of true baseline differences, which are important in what might become unintentionally a nonequivalent groups design. Whereas controlling statistically for individual subjects' means might unnecessarily remove some of the differential treatment effects, adjusting for individual intercepts will not have this undesirable influence.

Furthermore, many longitudinal studies use an assortment of covariates (clients' demographics, etc.) to adjust statistically for initial differences between treatment groups. In a previous substantive report on these same data, for example, Leff, *et al.* (1996) examined the effects of capitation on linear

trends for symptoms, social conflict and functioning using analysis of covariance (ANCOVA) and regression, controlling for sociodemographic and clinical variables. This procedure is used to make possible a nonequivalent groups design. However, because these particular covariates are only speculatively related to possible initial group differences and probably not even completely exhaustive thereof (although they might indeed represent valid predictors), they may not completely adjust or statistically control for the initial nonequivalence between groups. On the other hand, our use of individual subject intercepts on each of the criterion variables as covariates for the individual subject slopes, although admittedly not very explanatory in terms of the possible *causes* of the observed nonequivalence, is designed to completely remove any initial differences between groups (and between the constituent individuals) from the slopes. Furthermore, our use of the intercepts themselves as criterion variables in a prior regression model permits us to test the hypothesis of nonequivalence directly on the criterion variables of actual interest and not just on presumably influential demographics. Thus, it is possible that the previous report on these same data by Leff, *et al.* (1996) underestimated the initial nonequivalence between the groups.

Several intercept differences were found in the present analysis. Although these differences were partialled out in the analyses of the slopes, their presence highlights several methodological considerations in the present study, with implications for most other outcome studies. Examples of these findings will be presented to demonstrate these methodological concerns. Random assignment to groups is the cornerstone of any randomized clinical trial. In the present study, no intercept differences were found between the capitation group and the fee-for-service group which would seem to suggest effective randomization. However, an interaction between capitation and baseline functional status was found. Patients with personal skills assigned to the capitation condition had fewer psychological symptoms initially than patients with personal skills assigned to the fee-for-service condition, suggesting a problem with randomization. Additionally, patients without community skills assigned to the capitation condition had fewer psychological symptoms initially than patients without community skills assigned to the fee-for-service condition.

These findings clearly suggest problems in random assignment to treatment groups, i.e., methods of payment conditions, but an interaction between assignment to treatment groups and county was also found, which suggests another methodological problem. In Pima County, patients assigned to the fee-for-service condition had more psychological symptoms initially than patients assigned to the capitation condition. This pattern was reversed in Maricopa County: patients assigned to the capitation group had more psychological symptoms initially than patients assigned to the fee-for-service

group. It is impossible to know if this finding is due to differential biases in measurement, to variation in the relationships between symptoms and the blocking variables of diagnosis and functional level, or to differential biases in assignment of patients to groups at different sites. If this finding is the result of biases in measurement, however, it points to the need for better training of the persons who are rating the clients, in this instance the case manager. If these findings are due to variations in the relationships between the diagnosis and functional level blocking variables, then these relationships need to be better understood. Furthermore, evaluators of outcome studies often advocate the use of multiple sites in evaluating the effectiveness of an intervention. However, as the present findings demonstrate, steps need to be taken to ensure that measurement and procedures are standardized across sites. Issues of heterogeneity across treatment administration sites are relevant not only to the research methodology of program evaluation but to the formulation of treatment policy in the design of large-scale social interventions.

In addition to ensuring that measurement and procedures are standardized in a study with multiple sites, one must also consider whether there are differences in clients at different sites. Differences in clients may interact with how the treatment is administered, as well as affect the results. In the present study, patients in Pima County had better daily functioning initially than patients in Maricopa County. Thus, it appears that Pima County patients are healthier than their Maricopa County counterparts. Difference was also found between the sites within each county for patients assigned to the capitation groups. This difference may be indicative of a problem in randomization, however, rather than of a difference in clients at the two sites, since no difference was found between the sites overall.

In sum, the intercept is a more accurate covariate in attempting to control for initial differences than either the first observation or the mean. Examination of differences in intercept is also useful in that it indicates whether methodological problems, such as lack of randomization, measurement differences between raters, and heterogeneity of clients at various sites, exist in the study.

Differences in Slopes

The use of growth curve analysis in conjunction with general linear models allows one to examine multiple contrasts testing for differences in effects of method of payment based on site, location, and baseline status. While no main effects for capitation were found, interactions with capitation were found. Several of these contrasts provide information concerning whether capitation benefits certain groups more than others, as well as whether capitation is detrimental to certain groups. Some of the findings will be presented to illustrate this.

As stated previously, we found no main effect for method of payment, i.e., capitation *versus* fee-for-service care. An interaction between capitation, county, and baseline status was found, however. Patients in Pima County without functional skills receiving services under capitation experienced a decrease in psychological symptoms, while patients in Maricopa County without functional skills receiving services under capitation remained the same. This suggests that there may have been a difference in how the pilot programs were administered in the two counties, since differences due to initial status have been removed. In terms of evaluation it would be important in a case like this to find out what was different between the two counties in terms of program characteristics. For example, there may have been differences in staff qualifications or in how the program was implemented in each county. Documentation of program implementation was obtained (Leff, *et al.*, 1994); however, implementation differences between sites have not been fully analyzed.

Contrasts were also made based on baseline functional status. These contrasts provided information concerning whether initial assessment level helped to predict performance. In most of the contrasts, the higher the initial status, the greater the improvement in psychological symptoms and daily functioning, regardless of whether the patient received services under capitation. There were, however, instances of method of payment interacting with initial assessment status. For example, patients without functional skills who did not receive services under capitation experienced a dramatic increase in psychological symptoms, while patients without functional skills who received services under capitation improved considerably, as did patients with functional skills, although not as much. Patients with functional skills who did not receive services under capitation stayed about the same. A similar pattern was found for patients without community skills. Patients without community skills who did not receive services under capitation experienced a sharp increase in psychological symptoms. Although the treatment groups experienced a decrease in psychological symptoms as predicted, patients with community skills who did not receive services under capitation experienced the greatest improvement in psychological symptoms. These findings suggest while initial status is an important predictor of outcome, services under capitation are indeed effective for the more severely disturbed patients. They experience the greatest declines with the fee-for-service care. However, at the higher functioning level, in the presence of community skills, fee-for-service care is actually best. These interactions involving initial status have important implications for the care of the mentally ill. Leff, *et al.* (1996) found similar results using different analytic methods. That paper also offers several explanations for the differential effects of method of payment related to hypothesized financial incentives.

In sum, first, it is important to determine if there are differences in results between sites. The presence of differences indicates differences in program implementation that need to be followed up in order to determine which components of the program were most effective and which were least effective. This focus on initial differences will aid in designing future interventions. Second, it is also important to determine whether there is an interaction effect between client type and intervention. If an interaction is present, program design needs to take account of client status.

Difference in Residuals

The use of growth curve analysis allows one to examine any patterns suggested by the root-mean-squared residual, the scatter of observations for an individual around his or her own regression line. *A priori* contrasts can be used to predict this "error" term. The presence of significant contrasts would suggest more stability or variability in certain groups irrespective of any treatment effects or initial differences. The main finding was that there was more variability for patients in the capitation condition; patients in the fee-for-service condition were more stable. Apparently there were more ups and downs for patients receiving services under capitation, while there was less instability for patients receiving fee-for-service care. This effect was largely due to variability in Pima County patients receiving services under capitation. Maricopa County patients receiving services under capitation tended to be more stable. Differences in intraindividual variability between sites within each county were also found. Once again this highlights the possibility of differences in program implementation between the two counties and between sites within each county.

Differences due to an interaction between baseline functional status and capitation were also found. There was more intraindividual variability in daily functioning in patients without functional skills receiving fee-for-service care than patients without functional skills receiving services under capitation; however, the reverse was true for patients with functional skills. Patients with functional skills receiving services under capitation evidenced more intraindividual variation than patients receiving fee-for-service care. As discussed earlier in relation to effects of method of payment, it appears that there are different needs of the mentally ill based on their initial assessment status. These differences in needs have implications for program design.

The results discussed above indicate differences in program implementation between counties and sites within counties, suggesting the need for careful documentation of program implementation. McGrew, Bond, Dietzen, and Salyers (1992) have reported substantial differences in outcomes of treatment of severely mentally ill patients as a function of the integrity (Sechrest, West, Phillips, Redner, & Yeaton, 1979) of program implementation.

Another virtue of growth curve analysis that can be profitably exploited is that services which vary in amount over time for persons with severe mental illness can be represented as time-varying covariates in analyses of the effects of specific services (Gibbons, *et al.*, 1988). This would be a promising next step in applying these methods. The findings reported here are consistent in supporting the conclusion that needs of the mentally ill vary based on initial assessment status, a potentially vital consideration when interventions for the mentally ill are designed.

Finally, we would like to reiterate our belief that a meta-analytic approach to growth curve analyses represents a useful strategy in analyses of data in which systematic change is expected over time. In our view, the analyses reported here are not more complicated than would have been true of any reasonably satisfactory alternatives (e.g., Leff, *et al.*, 1996), and they are in many ways simpler since critical outcome constructs are reduced to a few easily comprehended parameters and require only commonly available and relatively user-friendly statistical software. This is an advantage for those longitudinal researchers who do not have either the resources or the inclination to invest time or money or both in the mastery of specialized software such as programs dedicated exclusively to the estimation of Hierarchical Linear Models. The meta-analytic approach to growth curve analysis is relatively accessible to users with average statistical training. Another advantage of our approach is its relative transparency, where every step taken is visible and the various intermediate numbers obtained along the way are directly interpretable. Many of the more sophisticated techniques available are rather opaque, meaning that you do not quite get to see just what is happening to your data. We believe that the insights afforded by the analyses of growth curves are provocative and meaningful and should be made available to a wider population of potential users. The meta-analytic approach to growth curve analysis is only one of several methods now available, each with its own merits and limitations, but we hope that it will at least provide a better gateway for more widespread application of this valuable analytic strategy for longitudinal data.

REFERENCES

- BARKHAM, M., REES, A., SHAPIRO, D. A., HARDY, G. E., STILES, W. B., & REYNOLDS, S. (1997) Dose-effect relations in time-limited psychotherapy for depression. *Journal of Clinical and Consulting Psychology*, 64, 927-935.
- BREKKE, J. S., LONG, J. D., NESBITT, N., & SOBEL, E. (1997) The impact of service characteristics on functional outcomes from community support programs for persons with schizophrenia: a growth curve analysis. *Journal of Clinical and Consulting Psychology*, 65, 464-475.
- BRYK, A. S., & RAUDENBUSH, S. W. (1992) *Hierarchical linear models: applications and data analysis methods*. Newbury Park, CA: Sage.
- COHEN, J., & COHEN, P. (1983) *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

- COLLINS, L. M., & HORN, J. L. (1991) *Best methods for the analysis of change: recent advances, unanswered questions, future directions*. Washington, DC: American Psychological Association.
- COOK, T. D., & CAMPBELL, D. T. (1979) *Quasi-experimentation: design and analysis issues in field settings*. Boston, MA: Houghton Mifflin.
- CUNNINGHAM, W. R. (1991) Issues in factorial invariance. In L. M. Collins & J. L. Horn (Eds.), *Best methods for the analysis of change: recent advances, unanswered questions, future directions*. Washington, DC: American Psychological Association. Pp. 106-125.
- FURMAN, W. M., & LUND, D. A. (1979) The assessment of patient needs: description of the level of care survey. *Journal of Psychiatric Treatment and Evaluation*, 1, 29-37.
- GIBBONS, R., HEDEKER, D., WATERNAUX, C., & DAVIS, J. (1988) Random regression models: a comprehensive approach to the analysis of longitudinal psychiatric data. *Psychopharmacology Bulletin*, 24, 438-443.
- HEDGES, L. V., & OLKIN, I. (1985) *Statistical methods for meta-analysis*. New York: Academic Press.
- LAIRD, N. M., & WARE, J. H. (1982) Random-effects models for longitudinal data. *Biometrics*, 38, 963-974.
- LEFF, H. S. (1992) Validation of a mental health service system model. Progress report submitted to the National Institute of Mental Health for Grant No. 5 R01 MH44878-02.
- LEFF, H. S., MULKERN, V. M., LIEBERMAN, M., & RAAB, B. (1994) The effects of capitation on service access, adequacy, and appropriateness. *Administration and Policy in Mental Health*, 21, 141-160.
- LEFF, H. S., MULKERN, V. M., LIEBERMAN, M., & RAAB, B. (1996) Outcome trends for severely mentally ill persons in capitated and case managed mental health programs. *Administration and Policy in Mental Health*, 24(1), 3-11.
- LEFF, H. S., MULKERN, V. M., RAAB, B., & WHITE, C. (1991) *Getting started: the Arizona Pilot Program Evaluation executive summary*. Cambridge, MA: Human Services Research Institute.
- MCARDLE, J. J. (1988) Dynamic but structural equation modeling of repeated measures data. In J. R. Nesselroade & R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology*. (2nd ed.) New York: Plenum. Pp. 561-614.
- MCARDLE, J. J., & ANDERSON, E. (1989) Latent growth models for research on aging. In J. E. Birren & K. W. Schaie (Eds.), *The handbook of the psychology of aging*. Vol. III. San Diego, CA: Academic Press. Pp. 21-44.
- MCARDLE, J. J., & HAMAGAMI, F. (1991) Modeling incomplete longitudinal and cross-sectional data using latent growth structural models. In L. M. Collins & J. L. Horn (Eds.), *Best methods for the analysis of change: recent advances, unanswered questions, future directions*. Washington, DC: American Psychological Association. Pp. 276-304.
- MCCAIN, L. J., & MCCLEARY, R. (1979) The statistical analysis of the simple interrupted time-series quasi-experiment. In T. D. Cook & D. T. Campbell (Eds.), *Quasi-experimentation: design and analysis issues for field settings*. Boston, MA: Houghton-Mifflin. Pp. 233-293.
- MCGREW, J. H., BOND, G. R., DIETZEN, L., & SALYERS, M. (1992) Measuring fidelity of implementation of a mental health program model. Paper presented at American Psychological Association Convention, Washington, DC, 17 August, 1992.
- MEREDITH, W. (1991) Latent variable models for studying differences and change. In L. M. Collins & J. L. Horn (Eds.), *Best methods for the analysis of change: recent advances, unanswered questions, future directions*. Washington, DC: American Psychological Association. Pp. 149-163.
- NESSELROADE, J. R. (1991) Interindividual differences in intraindividual change. In L. M. Collins & J. L. Horn (Eds.), *Best methods for the analysis of change: recent advances, unanswered questions, future directions*. Washington, DC: American Psychological Association. Pp. 92-105.
- PEDHAZUR, E. J. (1982) *Multiple regression in behavioral research: explanation and prediction*. New York: Holt, Rinehart & Winston.
- PETRINOVICH, L., & WIDAMAN, K. F. (1984) An evaluation of statistical strategies to analyze repeated-measures data. In H. V. S. Peeke & L. Petrinovich (Eds.), *Habituation, sensitization and behavior*. New York: Academic Press. Pp. 156-201.

- ROGOSA, D. R. (1989) A growth curve approach to the analysis of quantitative change. Invited presentation at "Best Methods for Analysis of Change" Conference, Los Angeles, October.
- ROGOSA, D. R., BRANDT, D., & ZIMOWSKI, M. (1982) A growth-curve approach to the measurement of change. *Psychological Bulletin*, 92, 726-748.
- ROGOSA, D. R., & SANER, S. (1994) Longitudinal data analysis examples with random coefficient models. Presentation at Rand Conference, October 1993. (Manuscript cited with permission)
- ROGOSA, D. R., & WILLETT, J. B. (1985) Understanding correlates of change by modeling individual differences in growth. *Psychometrika*, 50, 203-228.
- ROSENTHAL, R. (1969) Interpersonal expectations: effects of the experimenter's hypothesis. In R. Rosenthal & R. L. Rosnow (Eds.), *Artifact in behavioral research*. New York: Academic Press. Pp. 181-277.
- RYAN, C. S., SHERMAN, P. S., & BOGART, L. M. (1997) Patterns of services and consumer outcome in an intensive case management program. *Journal of Clinical and Consulting Psychology*, 65, 485-493.
- RYAN, C. S., SHERMAN, P. S., & JUDD, C. M. (1994) Accounting for case manager effects in the evaluation of mental health services. *Journal of Clinical and Consulting Psychology*, 62, 965-974.
- SAS INSTITUTE, INC. (1989) *SAS language and procedures: usage, Version 6, first edition*. Cary, NC: SAS Institute, Inc.
- SECHREST, L., & FIGUEROA, A. J. (1993) Program evaluation. *Annual Review of Psychology*, 44, 645-674.
- SECHREST, L., WEST, S. G., PHILLIPS, M., REDNER, R., & YEATON, W. H. (1979) Some neglected problems in evaluation research: strength and integrity of treatments. In L. Sechrest, et al. (Eds.), *Evaluation studies review annual*. Vol. IV. Beverly Hills, CA: Sage. Pp. 15-35.
- SHADISH, W. R., JR., COOK, T. D., & LEVITON, L. C. (1991) *Foundations of program evaluation: theories of practice*. Newbury Park, CA: Sage.
- SUTCLIFFE, J. P. (1980) On the relationship of reliability to statistical power. *Psychological Bulletin*, 88, 509-515.
- WIDAMAN, K. F. (1991) Qualitative transitions amid quantitative development: a challenge for measuring and representing change. In L. M. Collins & J. L. Horn (Eds.), *Best methods for the analysis of change: recent advances, unanswered questions, future directions*. Washington, DC: American Psychological Association. Pp. 204-217.
- WILLETT, J. B., & SAYER, A. G. (1994) Using covariance structure analysis to detect correlates and predictors of individual change over time. *Psychological Bulletin*, 166, 363-381.

Accepted August 14, 2000.