Clinical Practice as Natural Laboratory for Psychotherapy Research

A Guide to Case-Based Time-Series Analysis

Jeffrey J. Borckardt — *Medical University of South Carolina*
Michael R. Nash — *University of Tennessee*
Martin D. Murphy — *University of Akron*
Mark Moore — *Pennsylvania Hospital*
Darlene Shaw and Patrick O'Neil — *Medical University of South Carolina*

Both researchers and practitioners need to know more about how laboratory treatment protocols translate to real-world practice settings and how clinical innovations can be systematically tested and communicated to a skeptical scientific community. The single-case time-series study is well suited to opening a productive discourse between practice and laboratory. The appeal of case-based time-series studies, with multiple observations both before and after treatment, is that they enrich our design palette by providing the discipline another way to expand its empirical reach to practice settings and its subject matter to the contingencies of individual change. This article is a user's guide to conducting empirically respectable case-based time-series studies in a clinical practice or laboratory setting.

Keywords: time series, single-subject research, time-series analysis, psychotherapy research

Continuously tracking the symptom status of one (or a few) psychotherapy patients across baseline and intervention phases can potentially yield data sets well suited to revealing whether, when, and sometimes even why an intervention works. Peterson (2004) predicted, "Databases grounded in the actual experiences practitioners encounter will provide a descriptive foundation for a science that suits the nature we are trying to comprehend" (p. 205). Peterson is not alone in this sentiment. Barlow and Hersen (1984), Bergin and Strupp (1970), and Kazdin (1982, 1992) have long noted that the practitioner-generated case-based time-series design with baseline measurement fully qualifies as a true experiment and that it ought to stand alongside the more common group designs (e.g., the randomized controlled trial, or RCT) as a viable approach to expanding our knowledge about whether, how, and for whom psychotherapy works.

Many of the early breakthrough discoveries of psychological science were products of single-organism methodologies (Ebbinghaus, 1913; Fechner, 1889; Kohler, 1925; Pavlov, 1927; Skinner, 1938; Watson, 1925; for a review, see Morgan & Morgan, 2001). Skinner was a particularly strong advocate of single-organism time-series designs, probably because they allow for tracking what interested him most: when, how, and under what conditions new behavioral repertoires unfold in real time. In his view, the dominant large-*N* paradigm in psychology and its focus on group means actually obscured the anatomy of change.

Though the tradition of case-based time-series design with baseline measurement persists in the operant literature (R. R. Jones, Vaught, & Weinrott, 1978; Michael, 1974; Morgan & Morgan, 2001), it has languished in the long shadow cast by group methodologies for half a century. This is testimony to the robust yet flexible properties of group designs coupled with powerful statistical procedures singularly well suited to managing intersubject variability and questions of aggregate benefit. Two other factors may contribute to the neglect of time-series designs. First, the sheer prestige of group designs is so complete that the questions psychologists ask about psychotherapy outcome are almost always anchored to aggregate effect, the generic question being: Is the group mean of the criterion measure different for treated versus untreated subjects (or for different treatment conditions)? Hence Skinner's (1938) question about the anatomy of therapeutic change is rarely considered, let alone answered (but see Lambert, Hansen, & Finch, 2001; Price & Jones, 1998). Finally, the case-based time-series study in psychotherapy outcome research bears a special public relations burden of its own—its association with the field's early overreliance on unsubstantiated clinical anecdote.

Still, the call for empirically sturdy case studies survives and is now amplified. A number of researchers have

Jeffrey J. Borckardt, Darlene Shaw, and Patrick O'Neil, Department of Psychiatry and Behavioral Sciences, Medical University of South Carolina; Michael R. Nash, Department of Psychology, University of Tennessee; Martin D. Murphy, Department of Psychology, University of Akron; Mark Moore, Palliative Care Program, Pennsylvania Hospital, Philadelphia, PA.

Correspondence concerning this article should be addressed to Jeffrey J. Borckardt, Medical University of South Carolina, 67 President Street, IOP 5-North, 518, Charleston, SC 29425. E-mail: borckard@musc.edu

Vol. 63, No. 2, 77–95 DOI: 10.1037/0003-066X.63.2.77

Jeffrey J. Borckardt

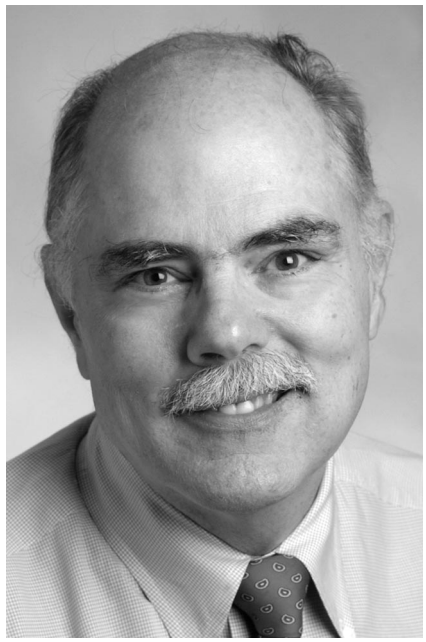
expressed an interest in whether and how laboratory-validated interventions translate to practice settings (Jacobson & Christensen, 1996; Westen & Bradley, 2005; Westen, Novotny, & Thompson-Brenner, 2004). The American Psychological Association's (APA's) Division 12 Task Force on Promotion and Dissemination of Psychological Procedures has explicitly recognized time-series designs as important methodological approaches that can fairly test treatment efficacy and/or effectiveness (Chambless & Ollendick, 2001). The APA Task Force on Evidence-Based Practice (2005) has endorsed systematic single-case studies as contributing to effective psychological practice. Westen and Bradley (2005) suggested that psychotherapy researchers "would do well to use clinical practice as a natural laboratory for identifying promising treatment approaches" (p. 267). The field seems to be recognizing that assaying aggregate effect is not the only empirical window researchers have on the nature of therapeutic change and that systematic observation of one or a few patients can be scientifically sound and instructive.

Unfortunately, there has been no upsurge of empirically grounded time-series case studies; many practitioners still despair over the relevance of psychotherapy research to practice; no therapy has been designated as efficacious on the weight of time-series data, as prescribed by APA task forces; and arguments about efficacy and effectiveness are framed almost exclusively in terms of group designs (Jacobson & Christensen, 1996; Kotkin, Daviet, & Gurin, 1996; Morrison, Bradley, & Westen, 2003; Nathan, Stuart, & Dolan, 2000; VandenBos, 1996; Westen & Morrison, 2001). Hence, despite clarion calls for its resurrection, the time-series design in psychotherapy outcome research lies dormant. Why?

We agree with Peterson's (2004) analysis that the enthronement of nomothetic group designs has "pushed every other approach downward, leaving case study despised or outcast entirely at the bottom." (p. 205). However, we do not believe the neglect of case-based research is *entirely* a matter of a recalcitrant science unwilling to entertain the notion of a carefully conducted case study. Part of the problem resides with practitioners. Even with sincere and immediate interest in what works for their patients, practitioners are sometimes intimidated by (or dismissive of) research and often unfamiliar with the case-based time-series options. This is unfortunate. Though for the typical practitioner a controlled large-group study is impractical, with some initiative and imagination the same practitioner can carry out one or more perfectly viable time-series studies.

Their appealing attributes aside, no time-series study or aggregation of time-series studies can provide the sharp-edged causal clarity of well-conducted RCTs for outcome evaluation. Nothing in this article alters the reality that group experimental designs (e.g., RCTs) are rightly the mainstay of our clinical science. Large-*N* experimental studies are unambiguously the designs of choice when psychologists ask questions regarding aggregate effect, especially when social policy matters are in the mix.

Neither does the time-series design with baseline measures define the only approach to evidence-based case study. The statistically derived and robust notion of "reliable and clinically significant change" (Jacobson, Roberts, Berns, & McGlinchey, 1999; Jacobson & Truax, 1991) has been influential in case studies where only one pre- and one postmeasure are available. Lambert and others have championed a patient-focused, bottom-up approach to outcome assessment, with emphasis on frequent assessment during therapy, statistically derived benchmarks for progress, and close examination of aggregate growth curves and dose



Michael R. Nash

Martin D. Murphy

effects (Barkham, Gilbert, Connell, Marshall, & Twigg, 2005; Evans, Margison, & Barkham, 1998; Haynes & O'Brien, 2000; Lambert, 2005; Lambert et al., 2001, 2002). In addition, E. E. Jones and his colleagues (E. E. Jones, Ghannam, Nigg, & Dyer, 1993; Price & Jones, 1998) have proposed an elegant paradigm for case studies in psychotherapy process research using a Q-set methodology. The fundamental appeal of case-based time-series studies as we describe them here, with multiple observations both before and after treatment, is that they enrich our design palette, providing the discipline another way to expand its empirical reach to practice settings and its subject matter to the contingencies of individual change.

This article is a user's guide to conducting case-based time-series studies in a practice setting. First, we offer a sampler of clinical research questions that can be addressed by case-based studies. Second, we construct a hypothetical case that illustrates the structure of a time-series project now being conducted in a university-based outpatient psychotherapy clinic. This case also familiarizes the reader with the data array of a time-series study. Third, we present two actual case studies, each carried out in a different outpatient setting. Fourth, we move to the logistics of how a time-series study is efficiently conducted in an applied setting. Finally, we provide a step-by-step description of simulation modeling analysis (SMA) for time-series data and how the practitioner can use freely available software to analyze his or her real-world clinical practice data (i.e., relatively short streams of time-series data). The use of SMA requires minimal statistical sophistication, and an Appendix is provided for further reference.

The Domain of Clinical Research Questions Addressed by Case-Based Time-Series Studies

A survey of the many varieties of time-series designs is beyond the scope of this article. There are a number of comprehensive descriptions of these powerful methodological tools (e.g., Barlow & Hersen, 1984, and Kazdin, 1982). We focus primarily on the simple A–B (pre–post) design because it is the most fundamental unit of inferential analysis across time-series designs and hence relevant to all of them. Further, the logistics of an outpatient setting are such that an A–B design (with follow-up when possible) is a good place for a seriously curious practitioner to begin.

Broadly speaking, time-series designs in psychotherapy can address two types of questions: questions of improvement (Does the patient get better following onset of treatment?) and questions of process change (How does change unfold during treatment?). On both counts, the time-series design can be a vehicle for practice-generated knowledge to inform laboratory protocol and for laboratory-generated knowledge to inform clinical practice.

Questions of Improvement: Looking for an Effect of Phase

Questions of improvement are essentially questions about whether onset of treatment is associated with improvement. Put another way, the generic time-series improvement question asks whether there is an effect of phase: Is there meaningful change in the patient's key symptoms from the pretreatment baseline condition (Phase A) to the treatment condition (Phase B)? If there is improvement, is it noteworthy? The critical comparison is between the level of symptom scores reported or observed during Phase A (i.e.,

Mark Moore

Darlene Shaw

before treatment onset) and the level of scores or observations during Phase B (after treatment onset). The size and direction of the difference on each outcome variable are then the grist for statistical analysis. Because the focus of a time-series improvement study is the comparison of symptom scores between two phases (typically pre- versus post-treatment), we refer to the analysis as a phase-effect analysis.

How Time-Series Improvement Studies Help Bridge the Gap Between Practice and the Laboratory

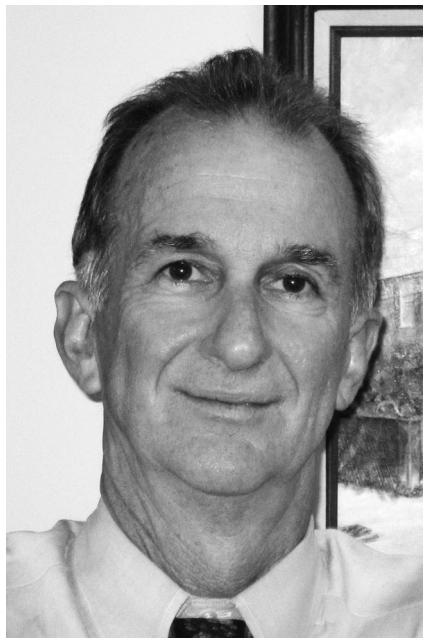
Laboratory research can be critically informed by carefully designed practitioner-conducted time-series improvement studies. For instance, practice-generated knowledge influences laboratory work when a practitioner reports a number of time-series cases testing whether an established empirically supported therapy (EST) for anxiety disorder might also benefit patients with impulse control disorder. By critically evaluating the practitioner's study, the laboratory researcher weighs whether the EST might have broader applications. This might then influence whether and how the laboratory researcher approaches future work on this technique. Similarly, when a practitioner conducts a time-series improvement study testing a new innovative therapy for hypertension, the laboratory researcher can assess the evidence and decide whether the new technique is promising enough to merit a closer look. Perhaps an RCT is in order. In both examples, evidence is brought to the table by the practitioner in a form that is epistemologically sound and therefore accessible and useful to the laboratory researcher.

Of course, there is nothing to stop a laboratory researcher from conducting her own time-series benefit studies. In doing so, she brings laboratory-generated knowledge to bear directly on clinical practice. For instance, she might test whether her EST for anorexia nervosa alone might also be associated with reduced symptoms among a series of anorectic patients comorbid for borderline personality disorder. In doing so, she brings to the table evidence that is singularly well suited for practitioners who comprehend the ideographic nature of the time-series format and who indeed frequently encounter patients with comorbid conditions (Westen & Bradley, 2005). In addition, the laboratory researcher has empirical findings on which she can base decisions as to whether this approach to comorbidity merits further investigation.

Questions of Process Change: Looking for Patterns of Change During Treatment

Questions of process change address how change unfolds over time and under what circumstances: the types of questions that interested Skinner (1938) the most. In contrast to improvement designs that require a comparison across two phases, process-change designs analyze change within one phase (usually the treatment phase). There are two types of process-change questions: univariate and multivariate. Univariate process change is addressed when one continuously tracks a single variable (e.g., symptom status) during treatment. One asks: Once in therapy, when does the patient begin to improve (latency)? At what pace does this improvement occur (slope)? These questions usually require only descriptive statistics.

Multivariate process change is addressed when one simultaneously tracks two or more processes during the course of treatment. One can then address mechanisms of change as well as sequencing. For instance, if during treatment one continuously tracks changes on a key symptom



Patrick O'Neil

(e.g., frequency of self-injurious behavior) against changes in the nature of the therapeutic interaction (e.g., status of rapport), one can ask: How is ongoing clinical improvement (a symptom variable) related to quality or frequency of ongoing events in-session (an intervention variable)? If symptom and intervention variables are related, in what order do they change? For example, are changes in rapport followed by symptom change in some systematic manner? Or does improvement (or decline) in symptom status precede change in rapport?

When one instead simultaneously tracks two key symptom variables, one can ask: Is there a sequence of improvement such that changes in Symptom A are followed by changes in Symptom B? If so, what is the lag? Whether tracking changes in symptom against changes in therapeutic interaction or symptom against symptom, the analysis is multivariate and must be sensitive to how multiple variables covary in real time. These analyses may likely involve cross-lagged correlations (discussed later).

How Time-Series Process-Change Studies Help Bridge the Gap Between Practice and the Laboratory

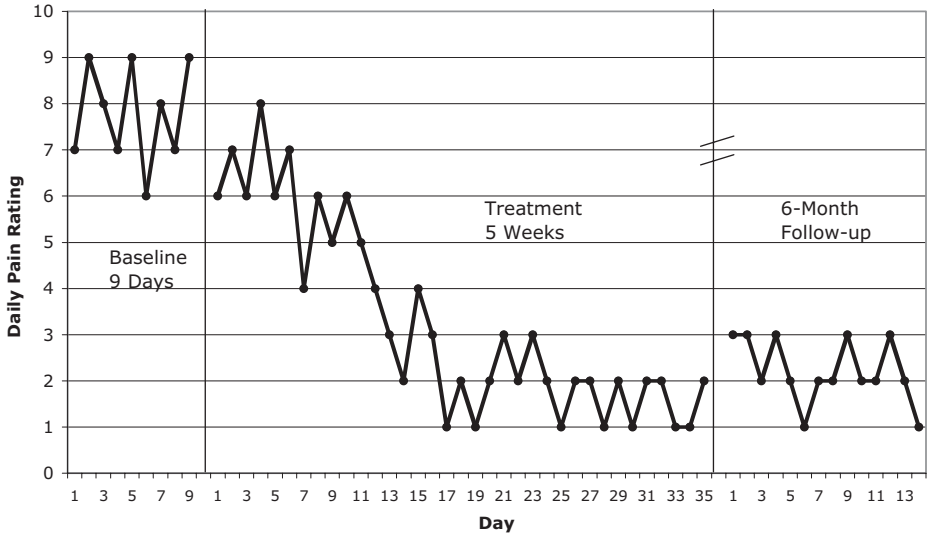
A practitioner might inform laboratory researchers about the change process when she conducts a number of time-series studies that track pain ratings, quality of life, and key intervention parameters among phantom limb pain patients before, during, and after psychological intervention. Laboratory researchers can then evaluate this evidence to look beyond benefit to questions of how, when, and under what specific therapeutic conditions phantom limb pain resolves. The findings might guide the timing and sequence of future laboratory interventions and offer leads to how the technique can be rendered maximally effective in as brief a time as possible. Similarly, if a theory predicts that an in-session technique or process (e.g., rapport, congruent empathy, exposure to feared stimuli, reinforcement) is mutative and should be followed by improvement, a practitioner-conducted time-series study can bring evidence to bear on the issue by simultaneously tracking the frequency or quality of these in-session events against the patient's day-to-day symptom ratings. There ought to be a relationship between symptom change and the purported mutative in-therapy event such that symptom change follows in-session events. Of course, similar studies conducted within a laboratory setting can inform clinicians about how change unfolds during the course of therapy, what aspects of the therapeutic environment are associated with benefit, and the extent of change to be expected.

A Real Practice-Based Time-Series Project Illustrated by an Imaginary Case

Beginning in 2002, the University of Tennessee Psychological Clinic has carried out empirically grounded case-based time-series studies with adult psychotherapy patients. The logistics of the Practice–Research Integrative Project (PRIP) evolved to better fit the pragmatic contours of the clinic setting (Nash, 2005). As such, the PRIP is a natural point of departure for this guide to conducting time-series studies in an outpatient setting.

The PRIP's structure is primarily (though not exclusively) a benefit design with baseline and treatment phases (see Figure 1). Follow-up at six months posttreatment is now part of the routine. Hence, these studies begin as A–B

Figure 1
Hypothetical Results Demonstrating a Typical Data Stream Encountered in Time-Series Studies: A Five-Week Intervention for Pain



designs with a generic protocol as follows: During a one- to two-week baseline phase and throughout the treatment phase, patients daily rate their symptom status on a general distress item (a Likert-type item ranging from 1 to 10) and two to three behavioral or self-report items well suited to the case formulation. The content of these latter items is crafted at intake by the therapist. Baseline observations for each patient typically number between 7 and 14. Treatment observations number at least 35. The total number of observations and the imbalance between baseline/treatment observations are typical of empirically grounded case studies in the literature (Center, Skiba, & Casey, 1985–1986; R. R. Jones et al., 1977). All patients complete the Outcome Questionnaire–45 (OQ-45; Lambert et al., 1996) at intake and monthly during treatment. In addition to this generic time-series protocol, therapists can add extra daily and monthly measures consistent with the treatment plan and research agenda.

Our hypothetical case illustrates in pure form the types of data streams generated by the PRIP protocol (see Figure 1). We posit an imaginary 53-year-old chronic pain patient treated over five sessions using a cognitive–behavioral approach coupled with self-hypnosis. The three measures tracked over time are as follows: daily pain intensity rating (from 1, *no pain,* to 10, *worst pain imaginable*); mood rating (from 1, *never bothered by depression,* to 10, *unrelenting and severe depression*), and the patient's reported level of distress (from 1, *none,* to 10, *severe*). For the purposes of this example, we show the pain rating alone.

Preliminary visual inspection of the hypothetical daily pain ratings (see Figure 1) might suggest improvement from baseline to treatment (baseline mean = 7.78; treatment mean = 3.29). Then again, even when seasoned judges use visual inspection of single-case data streams, they are prone to overestimate the effect of treatment (DeProspero & Cohen, 1979; Furlong & Wampold, 1982; R. R. Jones, Weinrott, & Vaught, 1978; Ottenbacher, 1993). This is especially true when, as in our case, observations are in principle, and in fact, not independent of one another (Borckardt, Murphy, Nash, & Shaw, 2004; R. R. Jones et al., 1978; Matyas & Greenwood, 1990; Robey, Schultz, Crawford, & Sinner, 1999). Matyas and Greenwood found visual inspection of these types of data streams to generate Type I error rates (false positives) of from 16% to 84%. Indeed, neither visual inspection nor conventional statistics are to be relied on for analyzing single-patient time-series studies (Robey et al., 1999) because such time-series data are autocorrelated.

What Is Autocorrelation?

Conventional parametric and nonparametric statistics assume that observations are independent. For instance, the result of a coin toss on Trial 1 does not influence the result on Trials 2 and 3, and so on. No matter how many times in a row "tails" is obtained, the probability that the next toss will be "heads" is unimpeachably still 50%. Hence, each observation (i.e., result of a coin toss) is independent. Similarly, in group designs, Subject 1's height is independent of Subject 2's height. Whether coin toss or height, one observation does not influence another. However, single-case time-series observations, such as the pain ratings in Figure 1, are in principle not independent. After all, the same person is generating the pain ratings. These data are in fact autocorrelated.

Simply put, a series of observations (as in the pain ratings in Figure 1) is said to be autocorrelated if the value of one observation depends (at least in part) on the value of one or more of the immediately preceding observations. Later observations are explained by earlier ones. Weather is autocorrelated. What the noon temperature will be on Wednesday is predicted by what the noon temperature was on Tuesday, and to a lesser extent what the noon temperature was on Monday or Sunday. Although the weather is certainly variable, how it changes from hour to hour, day to day, and season to season is to a degree lawful and structured, in a way that is *not* true when moving from one coin toss to the next. The stock market is autocorrelated. The value of the Dow Jones index at 2:00 p.m. is predicted by what it was at 1:00 p.m. if for no other reason than that the 2:00 p.m. value must proceed from the 1:00 p.m. value. Indeed, autocorrelation is an inevitable aspect of the periodicity, trending, and gradualism that one encounters regularly when tracking change over time in a single individual (weight loss, heart rate, tissue or psychological repair) or system (corporate earnings, birth rate). Autocorrelation is sometimes referred to as serial dependence. An early application of these notions to developmental psychology was described by Gottman and Ringland (1981).

How Is Autocorrelation Calculated?

How do we calculate the degree of autocorrelation? The primary focus of autocorrelation in the behavioral sciences is the Lag 1 correlation. The Lag 1 correlation is the degree to which an observation at Time K predicts the observation that comes immediately after it (at Time $K + 1$). Though calculation of autocorrelation is easily accomplished with statistical software, we believe that a one-time walk-through of its logic will inform the reader of what autocorrelation is conceptually.

Table 1 illustrates how a Lag 1 correlation could be calculated using data from our hypothetical case in Figure 1. To save space, we calculated the Lag 1 autocorrelation for the 35 pain ratings of the treatment phase only, but the principle can be applied to the entire data stream or to individual phase data streams. We expect that the pain ratings of our hypothetical patient in Figure 1 will be autocorrelated because all the ratings come from one person. The Lag 1 autocorrelation is the aggregate extent to which pain at Time 1 predicts pain at Time 2, pain at Time 2 predicts pain at Time 3, Pain at Time 3 predicts pain at Time 4, and so on. Hence, the Lag 1 correlation is simply the correlation between each data point and the point immediately following it. In Table 1, both Columns A and B depict the sequence of pain ratings as they occurred across time over the 35 observations of the treatment phase. The arrow from each pain rating in Column A leads to the pain rating that immediately follows it (in Column B). For example, the pain rating of 6 at Time 1 (in Column A) leads

Table 1
Calculating Lag 1 Autocorrelation of the 35 Treatment-Phase Pain Ratings in Figure 1: How Well Does Pain at Time K (Column A) Predict Pain at Time K + 1 (Column B)?

(Column A) Treatment-phase pain rating from Figure 1 at Time *K*	Predicts	(Column B) Pain rating at Time *K* + 1	(Column C) Pairs to be correlated: Time *K*, Time *K* + 1
6 (Time 1)	→	6 (Time 1)	—
7 (Time 2)	→	7 (Time 2)	6,7
6 (Time 3)	→	6 (Time 3)	7,6
8 (Time 4)	→	8 (Time 4)	6,8
6 (Time 5)	→	6 (Time 5)	8,6
7 (Time 6)	→	7 (Time 6)	6,7
4 (Time 7)	→	4 (Time 7)	7,4
6 (Time 8)	→	6 (Time 8)	4,6
5 (Time 9)	→	5 (Time 9)	6,5
6 (Time 10)	→	6 (Time 10)	5,6
5 (Time 11)	→	5 (Time 11)	6,5
4 (Time 12)	→	4 (Time 12)	5,4
3 (Time 13)	→	3 (Time 13)	4,3
2 (Time 14)	→	2 (Time 14)	3,2
4 (Time 15)	→	4 (Time 15)	2,4
3 (Time 16)	→	3 (Time 16)	4,3
1 (Time 17)	→	1 (Time 17)	3,1
2 (Time 18)	→	2 (Time 18)	1,2
1 (Time 19)	→	1 (Time 19)	2,1
2 (Time 20)	→	2 (Time 20)	1,2
3 (Time 21)	→	3 (Time 21)	2,3
2 (Time 22)	→	2 (Time 22)	3,2
3 (Time 23)	→	3 (Time 23)	2,3
2 (Time 24)	→	2 (Time 24)	3,2
1 (Time 25)	→	1 (Time 25)	2,1
2 (Time 26)	→	2 (Time 26)	1,2
2 (Time 27)	→	2 (Time 27)	2,2
1 (Time 28)	→	1 (Time 28)	2,1
2 (Time 29)	→	2 (Time 29)	1,2
1 (Time 30)	→	1 (Time 30)	2,1
2 (Time 31)	→	2 (Time 31)	1,2
2 (Time 32)	→	2 (Time 32)	2,2
1 (Time 33)	→	1 (Time 33)	2,1
1 (Time 34)	→	1 (Time 34)	1,1
2 (Time 35)	→	2 (Time 35)	1,2

Note. The Lag 1 autocorrelation = *r* (Time *K* × Time *K* + 1 in column C) = .81.

to the pain rating of 7 at Time 2 in Column B. Hence the first pair in the Lag 1 correlation calculation is 6, 7 (Column C). The Time 2 pain rating of 7 in Column A leads to the Time 3 pain rating of 6 in Column B; hence 7, 6 is entered in Column C. This process is followed through to the end of the data stream. The correlation of the pairs in Column C is the Lag 1 autocorrelation for the treatment phase, r(Lag 1) = .81. The pain ratings are strongly autocorrelated during the treatment phase. When we calculate the autocorrelation for the baseline and treatment phases taken together, the Lag 1 autocorrelation is .85.

Why Is Autocorrelation Important?

Autocorrelation is ubiquitous in behavioral data (Busk & Marascuilo, 1998; Sharpley & Alavosius, 1988). Under most circumstances, if the clinical investigator ignores autocorrelation with time-series data, he or she runs an unacceptably high risk of making a Type I error—that is, he or she infers that there is an effect of phase (from baseline to treatment) when in fact there is not. For this reason, all data sets in this article were analyzed using SMA for time-series, a statistical approach that accounts for autocorrelation and one that we describe later in detail.

Applying SMA to the Hypothetical Case

The key question for our hypothetical data set (see Figure 1) is an improvement (i.e., between-phases) question: Is the noted decrease in pain (from a mean of 7.78 during the baseline phase to a mean of 3.29 during the treatment phase) sufficiently improbable to justify our setting aside random variation in pain ratings as a viable explanation of the difference between pain in the baseline and treatment phases? In this case it is. Statistical analysis via SMA reveals that mere random variation of pain reports is an unlikely explanation of the phase difference from a baseline mean of 7.78 to the treatment mean of 3.29, $r(44) = -.69$, $p = .049$, even after controlling for autocorrelation. Further, at the six-month follow-up, the pain relief realized during treatment has remained fairly stable, with no discernible deterioration in relief: mean pain at treatment = 3.29; mean pain at follow-up = 2.21; $r(49) = -.26$, $p = .51$, *ns*. It is important to note that by comparing pain at baseline with pain at follow-up, we further confirm that at six months posttreatment the patient is experiencing less pain than he or she experienced prior to treatment: mean pain at baseline = 7.78; mean pain at follow-up = 2.21; $r(23) = -.96$, $p = .0002$. All of these analyses document a phase effect on a single variable (in this case, pain). A similar analysis could be applied to the daily distress measure and the mood measure. In any event, the patient is reporting less pain. Of course, whether pain receded because of the treatment itself or because of some other historical process is a matter to be addressed by subsequent group or time-series studies.

Real Cases

Below we briefly present two time-series case studies conducted in an outpatient psychotherapy setting. The first follows the university-based PRIP but was conducted in an Employee Assistance Program (EAP). Because it is primarily an improvement study, SMA was used to test for an effect of phase, controlling for autocorrelation. The second case is a time-series study conducted in a private practice setting. Its focus is a multivariate process-change relationship between two symptom variables, testing whether the pattern of changes on these variables conforms to what theories of therapeutic mechanism would predict. Both cases do double duty as real-world examples of the special (but not formidable) inferential and logistical challenges posed by time-series research in clinical practice.

Case 1—An Improvement Study in an EAP Setting: Looking for an Effect of Phase

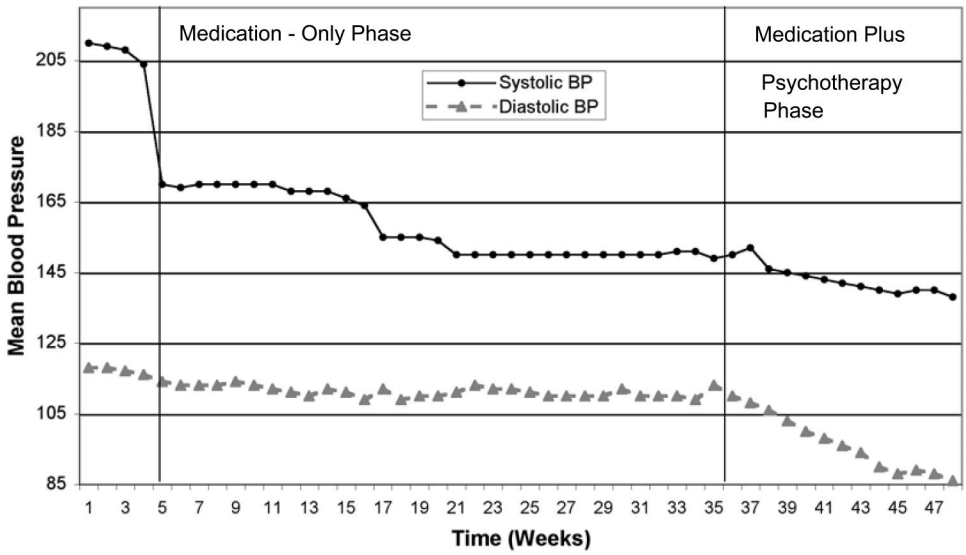
Reviews of the efficacy of psychosocial interventions for hypertension suggest weak to modest effects for cognitive–behavioral treatments (Eisenberg et al., 1993), relaxation therapies (Jacob, Chesney, Williams, Ding, & Shapiro, 1991; McGrady, 1996), education/training (Boulware et al., 2001), and stress management interventions (see Ebrahim & Smith, 1998). In this case, the patient's diligence inspired the therapist to test whether a psychotherapeutic intervention for hypertension would impact the patient's hypertension above and beyond the effects of pharmacological management.

A 42-year-old married, White, female, licensed practical nurse presented with a 10-month history of essential hypertension predominantly at work. During that time, the patient had meticulously kept a record of her self-monitored blood pressure taken daily in the work setting (hospital). She was consistent regarding recording times, equipment, and posture. Figure 2 illustrates that the pharmacological protocol applied during the nine months immediately prior to psychological intervention reduced the systolic blood pressure (SBP) from approximately 210 mmHg to a little less than 150 mmHg. However, during the same nine-month period, the patient's diastolic blood pressure (DBP) decreased very little, from approximately 120 mmHg to the still unacceptable range from 108 to 112 mmHg. The patient was referred for psychological intervention in hopes of augmenting the effect of medications.

The therapist (J.B.) instructed the patient to continue monitoring her blood pressure in the work setting. The ensuing 12-week therapeutic intervention was described by the therapist as incorporating insight-oriented and cognitive–behavioral approaches (for details, see Borckardt, 2001). The time-series design (depicted in Figure 2) tracked SBP and DBP across baseline, medication-alone, and medication-plus-psychotherapy phases, with sufficient precision to conduct a fair test of (a) whether the patient improved during psychological intervention, (b) whether the effect was statistically significant, and (c) whether the effect, if any, was on SBP, DBP, or both.

Analyses of the data in Figure 2 show that the reduction in SBP from the baseline condition to the medication-alone condition was statistically significant (Phase A, $M = 207.75$; Phase B, $M = 157.84$; $r = -.89$, $p < .001$); however, the effect on SBP of adding psychotherapy to the medication was not (Phase B, $M = 157.84$; Phase C, $M = 143.08$, $r = -.66$, $p = .20$). For DBP, the impact of medication was also statistically significant (Phase A, $M = 117.25$; Phase B, $M = 110.87$, $r = -.80$, $p < .001$); however, adding psychotherapy to medication enabled the patient to lower her DBP beyond that achieved by medication alone (Phase B, $M = 110.87$; Phase C, $M = 96.62$, $r = -.79$, $p = .03$). For this patient, the psychotherapy intervention was associated with a reduction of DBP (but not SBP) beyond that achieved with a variety of medications. Taken alone, these practice-based findings are not definitive. However, studies like these (carried out in a practice or a laboratory setting) might alert researchers to a promising intervention in this otherwise grim literature and might further inform them of what aspects of hypertension are most responsive to psychosocial intervention.

Figure 2
Case 1—Opportunistic Benefit Study: Mean Weekly Blood Pressure (BP) Readings (Taken at Work) Across Baseline, Medication-Only, and Medication-Plus-Psychotherapy Phases



Case 2—Private-Practice Setting: Looking for Patterns of Change During Treatment

In this example, the practitioner also found a significant phase effect, but we focus here on how to look beyond improvement to the process of change as it unfolds. Specifically, the practitioner asked: During a successful 31-session therapy with a depressed patient, does an increase in the extent of social engagement precede improvement of mood, or does improvement of mood precede an increase in social engagement? The answer hinges on a cross-lagged correlational analysis of data within the treatment phase.

A 42-year-old male patient presented in a private-practice setting with mixed anxiety and depressive symptoms. This symptom configuration had a 20-year history and had on one occasion required the patient's hospitalization. Although neither socially phobic nor agoraphobic, he generally avoided social gatherings. He secured employment enabling him to work from home or the office, as he chose. The patient's warm and intimate marital relationship was a notably bright spot in his life. Still, he struggled with substantial depression and could be alarmingly withdrawn socially. He had been prescribed several different antidepressants with only modest success. A life-long pattern of mood disorder and anxiety was clear. During initial intake, the therapist (M.N.) asked the patient to begin to record daily (a) the number of hours outside the house (social engagement) and (b) mood on a scale from 1 (*not depressed*) to 9 (*severely depressed*). These variables were chosen in collaboration with the patient as reasonably good indicators of "getting better." Each week, just prior to the psychotherapy session, the patient sent his rating sheet electronically to the therapist's office with date and numbers only (see Figure 3).

Improvement

The therapy was 31 sessions in duration. As Figure 3 illustrates, modest therapeutic gains were realized during therapy for daily mood and engagement ratings (the mean of the daily depression ratings for the first three weeks of therapy was 7.62, and for the last three weeks it was 2.67; the mean number of hours spent outside of the home each day during the first three weeks of therapy was 2.52, and during the last three weeks it was 5.24).

Process Change: The Cross-Lagged Correlation

Was the increase in social engagement followed by improvement in mood, or was the improvement in mood followed by greater social engagement? Perhaps there was no discernible pattern. Simultaneously graphing engagement and mood across sessions seems a promising approach to this problem. However, visual examination of the result (see Figure 3, which shows daily ratings of engagement and mood summed and averaged to weeks) is not immediately encouraging. The slopes of both mood and engagement track improvement, but is there any statistically discernible pattern such that change in one symptom leads or lags change in the other? And if so, by how many sessions does one lead or lag the other?

The cross-lagged correlation function is more revealing. The statistical details are described in the Appendix. Figure 4 summarizes the analysis. The horizontal axis represents the influence of one variable on the other expressed in lags. Lag 0 is the direct correlation of engagement at Time K with mood at Time K on a week-by-week basis (i.e., Time 1 engagement with Time 1 mood; Time 2 engagement with Time 2 mood, etc.). A lag of -5 is the correlation of engagement at Time K with mood five weeks earlier (Time $K - 5$). A lag of $+ 5$ is the correlation of the mood score with the engagement score five weeks hence. As Figure 4 illustrates, the largest cross-correlation is at the $+ 1$ lag, where mood precedes engagement by one week. The correlation coefficient is $-.82$ ($p < .00001$), significant with or without a Bonferroni correction for 11 comparisons (and when accounting for the influence of autocorrelation). Hence for this patient, about one week after his mood changed, his social behavior followed suit. This is an interesting, and to some degree unexpected, finding given current thinking that behavioral activation leads to improvement of mood among depressed patients (Hopko, Lejuez, Ruggiero, & Eiffert, 2003).

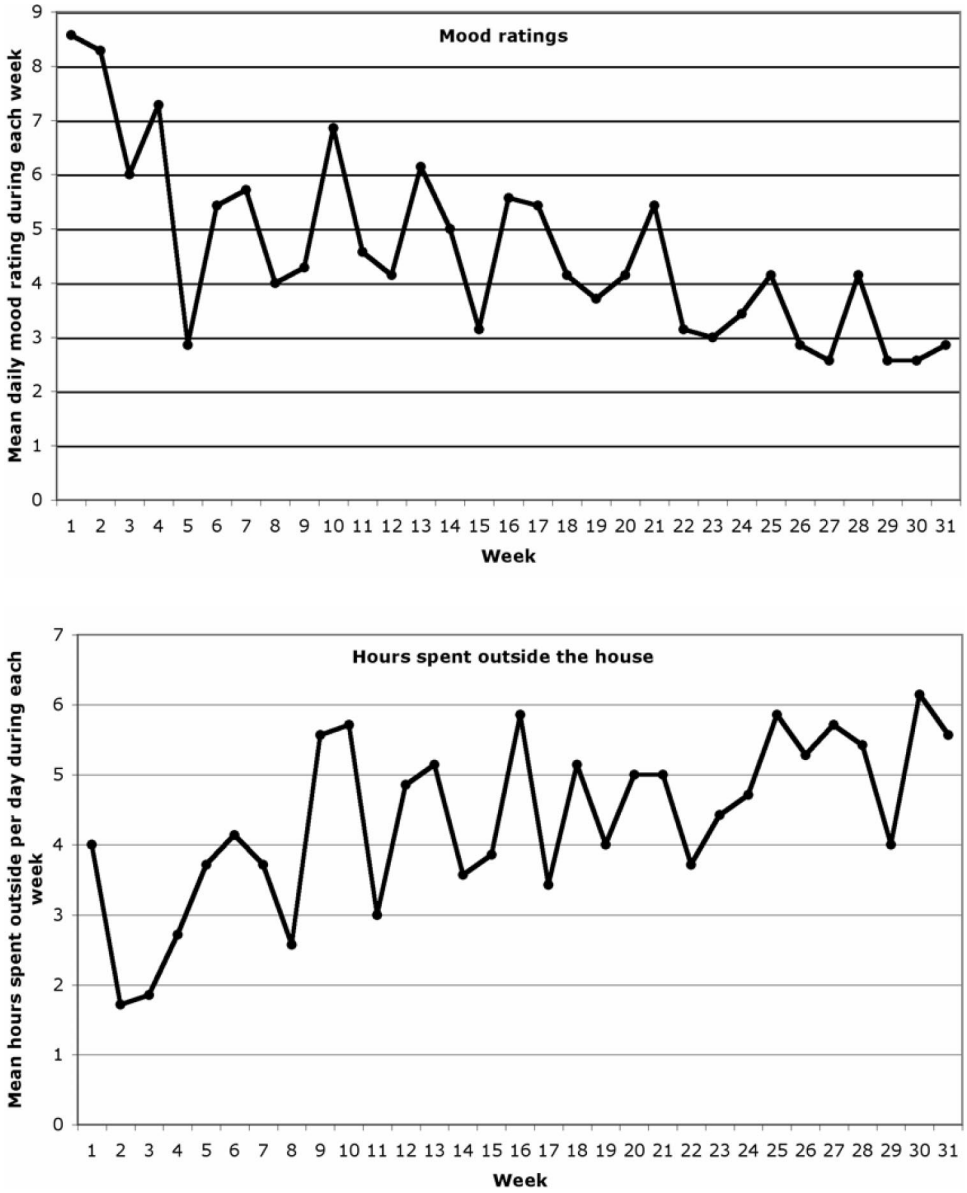
The evidence in this case is that mood leads activity level. Though alone one case is never definitive, it provides evidence that might move a laboratory researcher to revisit assumptions about the temporal sequence of behavioral activation and mood improvement in future designs. To the degree that this pattern is replicated, the need for revisiting theory becomes more urgent.

A Cautionary Note on Time and Cause

When we continuously track key symptoms and therapeutic parameters through the treatment phase, cross-lagged correlations of these data can reveal a great deal about how, to what degree, and in what order these processes are associated in time. This has implications for causal inference, but one must proceed carefully.

For instance, though statistically significant and sizable, the cross-lagged correlation finding in Case 2 ("change in mood" precedes "change in activity level") does not show that change in mood *causes* change in activity level. This would be an example of the post hoc fallacy (*post hoc ergo propter hoc*: after this, therefore because of this). Rather, if this finding were to be replicated in other studies, it would signal that the predominant behavioral activation model might need to be modified in some way. To understand why, one must first appreciate the relationship of time to cause. Here we are specifically addressing Aristotle's notion of efficient cause (i.e., triggering events) and not his notions of material, formal, or final causes (see Killeen, 2001; Killeen & Nash, 2003). Though it is a necessary condition to infer cause, merely showing that Event A preceded Event B does not prove that Event A caused Event B. For example, a hurricane (Event A) might precede the roof falling in (Event B), but that does not prove that the hurricane caused the roof to fall in (though it is suggestive). It is important to note that it is still true that an event cannot be caused by something that occurred *after* it happened: If the roof fell in (Event B)

Figure 3
Case 2—Pattern of Change: Sum of Mood Ratings and Hours Spent Outside of the House During a 31-Week Treatment



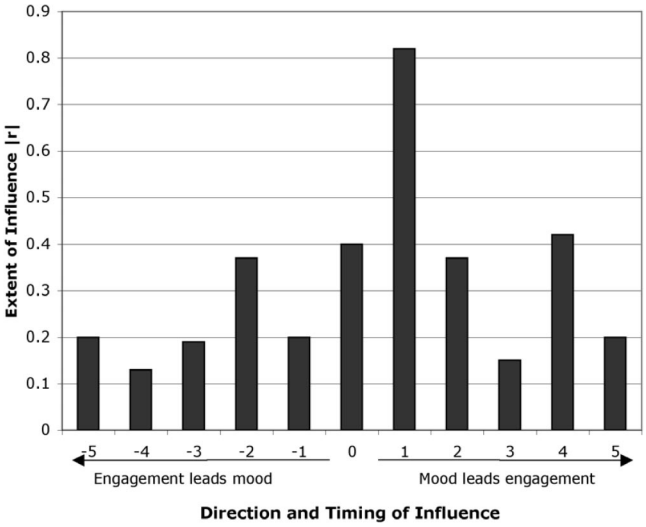
before the hurricane occurred (Event A), this would be inconsistent with the notion that the hurricane caused the roof to fall in. Hence, if a number of studies find that changes in mood precede changes in activity, this constitutes evidence that contradicts an uncomplicated linear causal model with behavioral activation driving mood change. Presumably, other models might then emerge accounting for this evidence. Our point here is that analysis of time-series data is singularly well suited to test whether the sequencing of the change process across time squares with theory. Conventional pre–post-follow-up group designs, though irreplaceable, do not allow one to test theory in this way.

Logistics: The Fundamentals

Creating a Data Stream

Tracking change over time is the most fundamental feature of single-case time-series outcome design. A patient is measured repeatedly on a number of outcome-related vari-

Figure 4
Case 2—Pattern of Change: Cross-Correlation Functioning Showing Directional and Temporal Relationship of Change in Mood With Change in Social Engagement During Therapy



ables across baseline, treatment, and follow-up phases. The outcome measures chosen are of course determined by the nature of the clinical problem, the opportunity for measurement, and the soundness of the measures themselves. At the onset, the clinical researcher must make critical decisions about the data stream.

Source and Content of the Observations

The source of the observations can be the patient, his or her significant others, and/or the therapist. The content of the measures might be symptoms (e.g., mood, anxiety, pain), social behaviors, physical status (weight, blood pressure), or medication usage. The case study is enriched when observations are secured across more than one source and more than one symptom cluster (Strupp, 1996). For instance, with a case involving treatment for pain, the patient might report level of pain on a daily basis and amount of prn (*pro re nata*, or as-needed) medication; the spouse might report the patient's level of activity and frequency of pain-related complaints.

Sample Evenly and Frequently

If change in the outcome variables is to be properly gauged, observations must be repeated evenly across time (e.g., daily, weekly). That is, the interval between measurements must be the same throughout the entire study; otherwise, statistical artifacts can occur. Whether the clinical focus is heart rate, blood pressure, hair pulling, itching, prn medication use, or self-mutilation, repeated observations sampled consistently over time and phase establish the topography of change. The number of observations for each phase can be different, but statistical analysis of intrasubject variability requires the *interval* of observation (e.g., daily, weekly) to be the same for all phases in the study. We have found daily measurement to be well tolerated by patients and well suited to the statistical requirements of the time-series analysis.

Because the clinical researcher is interested in knowing how these symptom data streams map against phases (e.g., baseline, treatment, follow-up), the total number of observations in the entire data stream and the number of observations in each phase are important. Statistically, the usual time-series study has about 10–20 total data points (Center, Skiba, & Casey, 1985–1986; R. R. Jones et al., 1977; Sharpley, 1987). SMA for time-series requires a minimum of 10–16 total observations in the data stream (i.e., 5–8 per phase).

Baseline Observations

In outpatient research, the number of baseline observations is at a premium. Understandably, patients object to delays in treatment. Statistically speaking, reasonable sensitivity and selectivity can be achieved with as few as 7–10 baseline observations. This is not difficult to realize. For instance, the PRIP intake clinician records a number of potential dependent variables in light of the presenting problem and proposed treatment plan. At the end of the intake interview, the patient is told that within a few days he or she will receive (a) a packet of customized daily rating sheets to track symptom status and (b) a telephone call from the assigned therapist to schedule a second appointment. Within two days after intake, three to four dependent variables are selected by the therapist, and customized response sheets are mailed to the patient.

The rating sheet (typically covering two to six days) is handed in by the patient upon arriving for the second appointment. During this meeting, the therapist and patient review the results of previous testing, elaborate on the nature and scope of the clinical problem, complete any further psychological testing if indicated, define the treatment plan, and schedule the first treatment session. Hence, when the patient returns for the next session (i.e., the first treatment session) he or she has completed seven daily ratings in addition to the two to six previously completed. In this manner, 9–13 baseline observations are realized before treatment begins. This process can be further streamlined in real time during the intake interview itself using software that enables the intake clinician to customize and print initial rating sheets for the patient to take home immediately after the intake session.

It must be remembered, however, that statistical power and internal validity are not the same thing. Although extended baselines are not strictly necessary to construct a statistical model of the data stream, stable baselines with many observations over a long period of time are *conceptually* preferable. When we have the luxury of long baselines, we feel more secure about attributing to the intervention clinical improvement observed after the onset of treatment. Still, the reality is that generating long baselines in a clinical setting is often not possible. For this

reason, a carefully conducted and well-documented history is an especially important feature of a time-series protocol. For instance, a patient reporting that she has not had sex with her husband since Christmas of the preceding year would inform a practitioner of the scope of the problem.

Supplementing With a Standardized Outcome Measure

Administration of a standardized outcome measure (e.g., the OQ-45, the Beck Depression Inventory, the Symptom Checklist–90) once at intake and periodically throughout treatment can enrich the time-series data (for a comprehensive review of outcome assessment measures, see Maruish, 2004). For example, the PRIP protocol requires administration of the OQ-45 (Lambert, Gregersen, & Burlingame, 2004) at intake and monthly thereafter. We chose the OQ-45 precisely because it provides cutoffs for reliable and clinically significant change (see Jacobson & Truax, 1991). Though statistical modeling of these data at the individual patient level is not viable given the small number of data points, if the trend of OQ-45 scores across time tracks reasonably well against the improvement indexed by daily observations, and meets the cutoff criteria for reliable and clinically significant change, an argument for benefit is enhanced. Supplemental assessment on a standardized measure also provides a common metric, allowing the clinical researcher to compare the extent and relevance of the patient's therapeutic gains with those of patients in other studies.

Maintaining the Data Stream

Patients return the single-page weekly rating sheets when they arrive for the therapy hour. This can be handled by a receptionist if one is available. For the PRIP project, the patient folds the weekly sheet and deposits it in a large locked box in the clinic waiting area that is clearly labeled "Response Sheets." The sheets themselves have no name—only a code number. This quickly becomes part of the routine of checking in. The receptionist keeps a record of whether the sheet is deposited. If a patient neglects to hand in a rating sheet, the therapist manages this as he or she would any other treatment-adherence issue. The timing of standardized measures (in the case of the PRIP, monthly administration of the OQ-45) is determined by a simple session count. For instance, every four weeks, PRIP patients are asked by the receptionist to complete the OQ-45 measure when they arrive for their therapy session. During the termination phase of therapy, patients are told that six months hence they will be contacted by mail for follow-up. In addition to the OQ-45, the follow-up packet includes rating sheets for 14 days.

The clinician (rather than the reception staff) could collect the daily rating sheets at the beginning of each session, and the data might even be integrated in some meaningful way with the therapy itself. However, in the case of the PRIP project, it was decided that it made sense from both training and clinical service perspectives to keep many of the logistics of the research agenda separate from the actual therapy. In real-world clinical settings, the individual clinician conducting the study should make decisions regarding data handling in consideration of potential clinical and logistical implications.

Ethical Considerations

Tracking the patient's symptom status, even if only in part, for the purpose of advancing science immediately requires attention to professional ethics beyond those routinely encountered in service delivery. These ethical matters involve informed consent, confidentiality, and the degree to which the research agenda compromises (or enhances) responsible service delivery. Some make a strong argument that ethical delivery of a treatment de facto requires attention to how the patient is responding (Association for Advancement of Behavior Therapy, 1977; Cone & Dalenberg, 2004; Hayes, 1981; Levy & Olson, 1979), thus justifying frequent assessment on service delivery considerations alone. Still, we assume that concessions to the research agenda are inevitable and must be addressed with the patient from the outset (Barrios, 1993; Bloom, Fischer, & Orme, 2003). We do this face to face during initial intake, addressing the issues as described below.

In our case studies, all assessment data are part of the patient's clinical chart. Hence the patient can expect confidentiality and accessibility as per professional ethical codes and HIPAA (Health Insurance Portability and Accountability Act of 1996) standards. At intake, the patient is informed of this in writing and is informed that, beyond good practice, one reason for our meticulous assessments is research: The patient's de-identified data might be included (possibly along with those of other patients) to help us learn more about how psychotherapy works. For the PRIP project, patients are told that normally the therapist does not see the data until therapy is finished. There is some flexibility in this, but whatever the arrangement is regarding the therapist's seeing the data, it is handled up front with the patient during the intake session. The patient is told that the treatment might be described in more detail in a scientific publication with his or her identity disguised. However, this would not occur unless or until the patient reads the report and gives consent (in writing) for us to share it with others. All of this is incorporated into our intake procedure, and patients can receive treatment without participating. We exclude patients presenting with emergent problems that might contraindicate even the minimal delays possible in our design (Kazdin, 1992).

Analysis of Time-Series Data

Data Fluctuation in Group Designs

Informally put, the generic outcome question for an RCT study asks, How viable is the notion that mere random sampling fluctuation (e.g., error variance) accounts for whatever benefit is observed in the treatment group relative to that of the control group? Scientific psychology has at its disposal a formidable array of parametric and nonparametric statistics specifically designed to detect nonrandom shifts in population parameters.

Data Fluctuation in Time-Series Studies

In a case-based time-series study, dispatching the "sampling fluctuation" explanation is more complex. This is because there is another source of variability peculiar to time-series designs: It is autocorrelation, or serial dependence. These are fluctuations that are due to monotonic trends, periodicity, or behavioral drifts in the data stream occurring across time (Suen, 1987). This lawful fluctuation is encountered in most other areas of natural science: meteorological shifts, economic recoveries, soil erosion, menses, and population genetics. As noted above, the presence of autocorrelation violates the fundamental assumption of conventional parametric and nonparametric statistics: independence of observations.

Autocorrelation as Nuisance and Raison d'Être

Observations are independent when each and every datum is its own unique source of information, unrelated to preceding or subsequent observations. In group research, this assumption is relatively secure. In time-series studies, where the patient's mood on Day 1 might very well partially determine mood on Day 2; and mood on Day 2 similarly becomes the point of departure for mood on Day 3, the data points are in a sense predicting each other. In other words there are not *really* as many observations as there *seem* to be because the observations are not independent of one another. For this reason, when conventional inferential group statistics (e.g., *t*, *F,* chi-square, and sign tests) are mistakenly applied to autocorrelated data streams, variability is underestimated; hence the effect/variability ratio is artifactually inflated. Spuriously high *t*s, *F*s, and *r*s, are generated, and researchers too often infer an effect when it is not justified (Hibbs, 1974; Sharpley & Alavosius, 1988).

The Nuisance

In practice, the incidence of autocorrelation in behavioral time-series data is generally viewed as sufficient to cause serious inferential bias if conventional statistics are used (Busk & Marascuilo, 1988; R. R. Jones et al., 1978; Matyas & Greenwood, 1990; Suen, 1987, but see Huitema, 1985, and Huitema & McKean, 1998; for reviews, see Franklin, Allison, & Gorman, 1996, and Robey et al., 1999). Further, it does not matter whether the autocorrelation coefficient is statistically significant. What matters is "the degree of distortion visited upon the *t* and *F* statistics when the autocorrelated data are analyzed via those procedures" (Sharpley & Alavosius, 1988, p. 246). For instance, whether it is significant or not, a calculated autocorrelation of .10 can inflate *t* and *F* values 110%–200% when the autocorrelation is .6. Note that the Lag 1 autocorrelations for the three cases presented in this article are as follows: Case 1—pain, $r = .81$; Case 2—SBP, $r = .87$, DBP, $r = .91$; Case 3—mood, $r = .42$, hours outside home, $r = .33$. At these levels of autocorrelation, Type I error rates can be significantly inflated. For this reason, time-series designs require special statistical treatment that adjusts for this problem.

As noted earlier, visual inspection of autocorrelated data streams, even by seasoned experts, yields low reliability and unacceptable Type I error rates. The same is true for nonparametric and modified parametric statistics (for reviews, see Franklin, Allison, & Gorman, 1996, and Robey et al., 1999). Multivariate software programs such as Autoregressive Integrated Moving Average Models (ARIMA), Hierarchical Linear Modeling (HLM), Interrupted Time-Series Analysis Correlational Analysis (ITSACORR), and autoregression essentially allow one to "model away" the autocorrelation and then test for significance (see also Gottman & Ringland, 1981; Price & Jones, 1998). These are powerful tools, but they require more observations per phase (at least 30–50) than are typically available in clinical work; and they statistically partial out serial dependence as though it were error.

The Promise

Though it is a statistical nuisance, by its very nature serial dependence reflects the momentum and gradualism of physiological, behavioral, and emotional repair. Because it is an index of serial dependence, the autocorrelation coefficient can reveal something about trends or fluctuations in symptoms before treatment and how these fluctuations shift during treatment. In a sense, autocorrelation is the natural subject matter of a clinical science. Whatever inferential statistic is applied to case-based time-series data streams, we believe it should approach autocorrelation not as noise that obscures change, but as music that attends it. Put differently, the preferred statistic gauges the occurrence of change while preserving its structure.

Simulation Modeling Analysis for Time-Series: Step by Step

SMA is a variant of bootstrapping methodologies that have been used to determine empirical significance levels across many kinds of data sets (see Wilcox, 2001). These approaches generally resample from known distributions to determine exact probabilities instead of probability estimates. For readers who are interested in a more detailed statistical treatment of SMA, please refer to the Appendix.

Below we walk the reader through SMA for time-series using a section of the DBP data from Case 1 (see Figure 2). For purposes of illustration we test whether there is a statistically significant effect on DBP for psychotherapy plus medication (Weeks 36–48) compared with medication alone (after the effects of the medication interventions appear to have stabilized; Weeks 21–35). There are three steps to this analysis, each enabled by the SMA computer software: (a) creation of the data stream array, (b) determination of effect size and autocorrelation, and (c) simulation modeling with generation of significance criteria.

Step 1: Data stream array. Table 2 illustrates the structure and logic of the data array. In the first column are the weeks, beginning with Week 21 (stabilized medication-only effect), through Week 35 (just before psychotherapy was added), and from Week 36 to Week 48 (during the medication-plus-psychotherapy phase). Column 2 records the mean DBP for the given week. The third column indicates whether the weekly mean DBP is during

Table 2
Structure and Logic of the Time-Series Data Array Using an Example of Weekly Diastolic Blood Pressure Before and After Onset of a Multimodal Psychotherapy (From Case 1 in Figure 2)

Week	Diastolic blood pressure (mmHg)	Phase vector dummy coding (0 = baseline, 1 = treatment)	Phase
21	111	0	
22	113	0	
23	112	0	
24	112	0	
25	111	0	
26	110	0	
27	110	0	Baseline phase
28	110	0	A, *n* = 15
29	110	0	
30	112	0	
31	110	0	
32	110	0	
33	110	0	
34	109	0	Onset of
35	113	0	treatment
36	110	1	
37	108	1	
38	106	1	
39	103	1	
40	100	1	Treatment
41	98	1	Phase B,
42	96	1	*n* = 13
43	94	1	
44	90	1	
45	88	1	
46	89	1	
47	88	1	
48	86	1	

Note. Phase A autocorrelation estimate = .17; Phase B autocorrelation estimate = .85; correlation of dependent variable and phase vector = −.79.

the baseline pharmacology-alone phase (coded 0) or during the pharmacology-plus-psychotherapy phase (coded 1). There are 15 observations in the baseline phase and 13 in the psychotherapy treatment phase.

Step 2: Determination of effect size and parameters for simulation modeling. To gauge the effect size for psychotherapy, the computer calculates the Pearson correlation between DBP and the phase dummy coding. The effect size in our example is −.79. In other words, DBP during the psychotherapy treatment phase is considerably lower than during the baseline medication phase. But we know that autocorrelation can create large effect sizes (i.e., high correlations between phase and DBP in our case) that might look significant but are not. To correct for this, we must know the extent of autocorrelation (AR) in the DBP measures over time. This is calculated by the SMA software via the formula shown in Step 2 of the Appendix. In our example, the autocorrelation, *r*(Lag 1), for the baseline phase of our observed data stream is .17, and the autocorrelation for the medication-plus-psychotherapy phase is .85. Hence, whatever correction there must be, it will be considerable.

Step 3: Simulation modeling with generation of significance criteria. SMA utilizes four parameters of the observed data stream—the autocorrelation estimate from the baseline phase, or Phase A (AR_a); the autocorrelation from the treatment phase, or Phase B (AR_b); the number of observations in Phase A (N_a); and the number of observations in Phase B (N_b)—to generate thousands of simulated data streams drawn randomly from a known null distribution of data streams (random normal with no programmed effects) all of which have the same autocorrelation and number of observations as the original observed data stream. When these thousands of data sets are evaluated, the user can determine how likely different effect sizes are given the specified levels of autocorrelation and N (when no effects are actually present). Note that this is virtually the same approach as Monte Carlo analysis but with specific values for the parameters of interest (phase N and autocorrelation values). In our example, all 10,000 simulated data streams have 15 baseline observations, 13 treatment observations, a baseline AR_a of .17, and a treatment-phase AR_b of .85. Remember that they are drawn from a null distribution of data streams that has no programmed effect (i.e., the mean effect size in this distribution is 0). Of course, there will be variability around zero, and this is what interests us. For each of these simulated data streams, the Pearson *r* value (i.e., treatment effect size) is calculated in the same fashion as it was for the observed patient data stream. A table is generated, as in Table 3, that gives the probability that a given effect size (an *r* of −.79 in our example) will occur by chance in a null distribution of data streams with AR_a = .17, AR_b = .85, N_a = 15, and N_b = 13. If we scroll down Table 3 to our observed treatment effect in our observed data stream (r = −.79), we find that the probability of obtaining an effect size of −.79 (or larger) in a distribution of data streams which is in fact null is .03. This probability then provides an empirical estimate of the probability of such an observed value occurring by chance.

As illustrated in Table A1 in the Appendix, SMA provides clinical researchers with an attractive method for evaluating the statistical significance of between-phase changes in data streams typical of most case-based clinical research: a total number of observations less than 30, an autocorrelation between .2 and .8, and moderate to large effect sizes (Center et al., 1985–1986; R. R. Jones et al., 1977; Sharpley, 1987). Note that when more data are available (30+ data points per phase), more conventional time-series analytic techniques are preferred (e.g., ARIMA, HLM, autoregression). However, with short data streams (between 5 and 15 data points per phase), SMA delivers substantially more power than conventional statistics (including several time-series-specific analytic approaches such as ITSACORR, Crosbie, 1993; see Table A1 in the Appendix), and its selectivity is acceptable. A free, user-

Table 3
Empirical Probability Table Generated via the Simulation Modeling Approach That Provides the User With the Probability of Discovering Various Effects (Correlations Between the Dependent Measure of Interest and the Baseline/Treatment Vector) Among Random Data With Properties (N and Autocorrelation) Similar to Those of the Data in Question

Absolute value of Pearson *r* (|*r*|)	Probability (*p*) of attaining |*r*| given *N* and autocorrelation	Absolute value of Pearson *r* (|*r*|)	Probability (*p*) of attaining |*r*| given *N* and autocorrelation	Absolute value of Pearson *r* (|*r*|)	Probability (*p*) of attaining |*r*| given *N* and autocorrelation
.00	1.000	.34	.538	.68	.117
.01	.987	.35	.524	.69	.106
.02	.977	.36	.512	.70	.097
.03	.963	.37	.502	.71	.089
.04	.947	.38	.492	.72	.081
.05	.935	.39	.478	.73	.073
.06	.924	.40	.462	.74	.064
.07	.910	.41	.447	.75	.056
.08	.897	.42	.433	.76	.047
.09	.885	.43	.419	.77	.041
.10	.874	.44	.406	.78	.035
.11	.858	.45	.393	.79	.030
.12	.846	.46	.379	.80	.036
.13	.832	.47	.367	.81	.022
.14	.819	.48	.350	.82	.017
.15	.802	.49	.338	.83	.014
.16	.786	.50	.326	.84	.011
.17	.773	.51	.314	.85	.009
.18	.762	.52	.303	.86	.006
.19	.747	.53	.293	.87	.005
.20	.734	.54	.279	.88	.003
.21	.720	.55	.267	.89	.001
.22	.706	.56	.255	.90	.001
.23	.694	.57	.243	.91	.000
.24	.681	.58	.228	.92	.000
.25	.667	.59	.217	.93	.000
.26	.652	.60	.205	.94	.000
.27	.639	.61	.196	.95	.000
.28	.625	.62	.186	.96	.000
.29	.611	.63	.174	.97	.000
.30	.598	.64	.160	.98	.000
.31	.582	.65	.148	.99	.000
.32	.570	.66	.138	1.00	.000
.33	.555	.67	.128		

Note. Data properties: Phase A $N = 15$; Phase B $N = 13$; Phase A autocorrelation = .17; Phase B autocorrelation = .85. Correlation between actual and phase vector: $r = -.79$, $p = .030$.

friendly stand-alone computer program has been developed to run SMA for Windows and Macintosh platforms. It can be downloaded at http://clinicalresearcher.org.

Conclusions

Case-based time-series designs will not dissolve the formidable epistemological gap between practice and research, but their use can help bring the two disciplines within shouting distance of each other on a more regular basis. As with many things in life, enhanced communication pivots on compromise by both parties. For their part, practitioners must concede that replicable systematic observation is a necessary requirement of evidence; in turn, researchers must concede (indeed rediscover) that carefully conducted ideographic studies can yield empirically sound findings about therapeutic change. In some cases, the evidence can be obtained in no other way.

For some, these concessions will not be forthcoming. Still, a robust clinical science requires an ongoing productive discourse between a critical mass of researchers and practitioners. Herein lies the twofold promise of case-based time-series designs. First, their careful use enables practitioners to make contributions that are fully congruent with the evidence-driven ethos of scientific discourse. By rising

above mere clinical anecdote, practitioners earn a more prominent and respected voice on matters of theory, research, policy, and training. The clinical setting can indeed become the natural laboratory envisioned by Westen and Bradley (2005) and Peterson (2004). Second, time-series designs yield findings especially pertinent to how therapeutic change unfolds, not in the aggregate, but individually. Though almost entirely neglected by contemporary investigators, single-subject research of this kind has a luminous and storied lineage in experimental and clinical psychology. By harnessing time-series designs alongside group experimental methodologies, psychologists will accelerate the progress we are making in understanding the structure and mechanism of therapeutic change.

REFERENCES

APA Presidential Task Force on Evidence-Based Practice. (2006). Evidence-based practice in psychology. *American Psychologist, 61,* 271–285.

Association for Advancement of Behavior Therapy. (1977). Ethical issues for human services. *Behavior Therapy, 8,* 763–764.

Barkham, M., Gilbert, N., Connell, J., Marshall, C., & Twigg, E. (2005). Suitability and utility of the CORE-OM and CORE-A for assessing severity of presenting problems in psychological therapy services based in primary and secondary care settings. *British Journal of Psychiatry, 186,* 239–246.

Barlow, D. H., & Hersen, M. (1984). *Single case experimental designs.* New York: Pergamon Press.

Barrios, B. A. (1993). Direct observation. In T. H. Ollendick & M. Hersen (Eds.), *Handbook of child and adolescent assessment* (pp. 140–164). Boston: Allyn & Bacon.

Bergin, A. E., & Strupp, H. H. (1970). New directions in psychotherapy research. *Journal of Abnormal Psychology, 76,* 13–26.

Bloom, M., Fischer, J., & Orme, J. G. (2003). *Evaluating practice: Guidelines for the accountable professional* (4th ed.). Boston: Allyn & Bacon.

Borckardt, J. J. (2001). Case study demonstrating the efficacy of a multimodal psychotherapeutic intervention for hypertension. *International Journal of Clinical and Experimental Hypnosis, 50,* 189–201.

Borckardt, J. J., Murphy, M. D., Nash, M. R., & Shaw, D. (2004). An empirical examination of visual analysis procedures for clinical practice evaluation. *Journal of Social Service Research, 30*(3), 55–73.

Boulware, L. E., Daumit, G. L., Frick, K. D., Minkovitz, C. S., Lawrence, R. S., & Power, N. R. (2001). An evidence-based review of patient-centered behavioral interventions for hypertension. *American Journal of Preventive Medicine, 21,* 221–232.

Busk, P. L., & Marascuilo, R. C. (1988). Autocorrelation in single-subject research: A counter-argument to the myth of no autocorrelation. *Behavioral Assessment, 10,* 229–242.

Center, B. A., Skiba, R. J., & Casey, A. (1985–1986, Winter). A methodology for the quantitative synthesis of intra-subject design research. *Journal of Special Education, 19,* 387–400.

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

Cone, J. D., & Dalenberg, C. J. (2004). Ethics concerns in outcome assessments. In M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcome assessment* (3rd ed., Vol. 1, pp. 335–368). Mahwah, NJ: Erlbaum.

Crosbie, J. (1993). Interrupted time-series analysis with brief single-subject data. *Journal of Consulting and Clinical Psychology, 61,* 966–974.

DeProspero, W., & Cohen, S. (1979). Inconsistent visual analysis of intrasubject data. *Journal of Applied Behavior Analysis, 12,* 573–579.

Ebbinghaus, H. (1913). *Memory* (H. Ruyer & C. E. Bussenius, Trans.). New York: Teachers College, Columbia University.

Ebrahim, S., & Smith, G. D. (1998). Lowering blood pressure: A systematic review of sustained effects of nonpharmacological interventions. *Journal of Public Health and Medicine, 20,* 441–448.

Eisenberg, D. M., Delbanco, T. L., Berkey, C. S., Kaptchuk, T. J., Kupelnick, B., Kuhl, J. & Chalmers, T. C. (1993). Cognitive behavioral techniques for hypertension: Are they effective? *Annals of Behavioral Medicine, 118,* 964–972.

Evans, C., Margison, F., & Barkham, M. (1998). The contributions of reliable and clinically significant change methods to evidence-based mental health. *Evidence-Based Mental Health, 1,* 70–72.

Fechner, G. T. (1889). *Elemente der Psychophysik* [Elements of psychophysics]. Leipzig, Germany: Breitkopf & Hartel.

Franklin, R. D., Allison, D. B., & Gorman, B. S. (1996). *Design and analysis of single-case research.* Mahwah, NJ: Erlbaum.

Furlong, M. J., & Wampold, B. E. (1982). Intervention effects and relative variation as dimensions on experts' use of visual inference. *Journal of Applied Behavior Analysis, 15,* 415–421.

Gottman, J. M., & Ringland, J. T. (1981). The analysis of dominance and bidirectionality in social development. *Child Development, 52,* 393–412.

Hayes, S. C. (1981). Single case experimental design and empirical clinical practice. *Journal of Consulting and Clinical Psychology, 49,* 193–211.

Haynes, S. N., & O'Brien, W. H. (2000). *Principles and practice of behavioral assessment.* New York: Kluwer Academic/Plenum.

Hibbs, D. A. (1974). Problems of statistical estimation and causal inference in time-series regression models. In H. L. Costner (Ed.), *Sociological methodology, 1973–74* (pp. 252–307). San Francisco: Jossey-Bass.

Hopko, D. R., Lejuez, C. W., Ruggiero, K. J., & Eiffert, G. H. (2003). Contemporary behavioral activation treatments for depression: Procedures, principles, progress. *Clinical Psychology Review, 23,* 699–717.

Huitema, B. E. (1985). Autocorrelation in applied behavior analysis: A myth. *Behavioral Assessment, 7,* 107–118.

Huitema, B. E., & McKean, J. W. (1998). Irrelevant autocorrelation in least-squares intervention models. *Psychological Methods, 3,* 104–116.

Jacob, R. G., Chesney, M. A., Williams, D. M., Ding, Y., & Shapiro, A. P. (1991). Relaxation therapy for hypertension: Design effects and treatment. *Annals of Behavioral Medicine, 13,* 5–17.

Jacobson, N. S., & Christensen, A. (1996). Studying the effectiveness of psychotherapy: How well can clinical trials do the job? *American Psychologist, 51,* 1031–1039.

Jacobson, N. S., Roberts, L. J., Berns, S. B., & McGlinchey, J. B. (1999). Methods for defining and determining the clinical significance of treatment effects: Description, application, and alternative. *Journal of Consulting and Clinical Psychology, 67,* 300–307.

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59,* 12–19.

Jones, E. E., Ghannam, J., Nigg, J. T., & Dyer, J. P. (1993). A paradigm for single-case research: The time series study of a long-term psychotherapy for depression. *Journal of Consulting and Clinical Psychology, 61,* 381–394.

Jones, R. R., Vaught, R. S., & Weinrott, M. R. (1977). Time-series analysis in operant research. *Journal of Applied Behavior Analysis, 10,* 151–166.

Jones, R. R., Weinrott, M. R., & Vaught, R. S. (1978). Effects of serial dependency on the agreement between visual and statistical inference. *Journal of Applied Behavior Analysis, 11,* 277–283.

Kazdin, A. E. (1982). *Single-case research designs: Methods for clinical and applied settings.* New York: Oxford University Press.

Kazdin, A. E. (1992). *Research design in clinical psychology* (2nd ed.). Boston: Allyn & Bacon.

Killeen, P. R. (2001). The four causes of behavior. *Current Directions in Psychological Science, 10,* 136–140.

Killeen, P. R., & Nash, M. R. (2003). The four causes of hypnosis. *International Journal of Clinical and Experimental Hypnosis, 51,* 195–231.

Kohler, W. (1925). *The mentality of apes.* New York: Harcourt.

Kotkin, M., Daviet, C., & Gurin, J. (1996). The *Consumer Reports* mental health survey. *American Psychologist, 51,* 1080–1082.

Lambert, M. J. (2005). Emerging methods for providing clinicians with

timely feedback on treatment effectiveness: An introduction. *Journal of Clinical Psychology, 61,* 141–144.
Lambert, M. J., Gregersen, A. T., & Burlingame, G. M. (2004). The Outcome Questionnaire-45. In M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcome assessment* (3rd ed., pp. 191–234). Mahwah, NJ: Erlbaum.
Lambert, M. J., Hansen, N. B., & Finch, A. E. (2001). Patient-focused research: Using patient outcome data to enhance treatment effects. *Journal of Consulting and Clinical Psychology, 69,* 159–172.
Lambert, M. J., Hansen, N. B., Umphress, V., Lunnen, K., Okiishi, J., Burlingame, G. M., et al. (1996). *Administration and scoring manual for the OQ-45.* Stevenson, MD: Professional Credentialing Services.
Lambert, M. J., Whipple, J. L., Bishop, M. J., Vermeersch, D. A., Gray, G. V., & Finch, A. E. (2002). Comparison of empirically-derived and rationally-derived methods for identifying patients at risk for treatment failure. *Clinical Psychology and Psychotherapy, 9,* 149–164.
Levy, R. L., & Olson, D. G. (1979). The single subject methodology in clinical practice: An overview. *Journal of Social Service Research, 3,* 25–49.
Maruish, M. E. (2004). *The use of psychological testing for treatment planning and outcomes assessment.* Mahwah, NJ: Erlbaum.
Matyas, T. A., & Greenwood, K. M. (1990). Visual analysis of single-case time series: Effects of variability, serial dependence, and magnitude of intervention of effects. *Journal of Applied Behavior Analysis, 23,* 213–224.
McGrady, A. (1996). Good news—Bad press: Applied psychophysiology in cardiovascular disorders. *Biofeedback and Self-Regulation, 21,* 335–346.
Michael, J. (1974). Statistical inference for individual organism research: Mixed blessing or curse? *Journal of Applied Behavior Analysis, 7,* 647–653.
Morgan, D. L., & Morgan, R. K. (2001). Single-participant research design: Bringing science to managed care. *American Psychologist, 56,* 119–127.
Morrison, K. H., Bradley, R., & Westen, D. (2003). The external validity of controlled clinical trials of psychotherapy for depression and anxiety: A naturalistic study. *Psychology & Psychotherapy: Theory, Research & Practice, 76(2*), 109–132.
Nash, M. R. (2005). *Practice–Research Integrative Project.* Unpublished manuscript, University of Tennessee, Knoxville.
Nathan, P. E., Stuart, S. P., & Dolan, S. L. (2000). Research on psychotherapy efficacy and effectiveness: Between Scylla and Charybdis? *Psychological Bulletin, 126,* 964–981.
Ottenbacher, K. J. (1993). Interrater agreement of visual analysis in single-subject decisions: Quantitative review and analysis. *American Journal of Mental Retardation, 98,* 135–142.
Pavlov, I. P. (1927). *Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex* (G. V. Anrep, Trans.). New York: Oxford University Press.
Peterson, D. R. (2004). Science, scientism, and professional responsibility. *Clinical Psychology: Science and Practice, 11,* 196–210.
Price, P. B., & Jones, E. E. (1998). Examining the alliance using the psychotherapy process Q-set. *Psychotherapy, 35,* 392–404.
Robey, R. R., Schultz, M. C., Crawford, A. B., & Sinner, C. A. (1999). Single-subject clinical-outcome research: Designs, data, effect sizes, and analyses. *Aphasiology, 13,* 445–473.
Sharpley, C. F. (1987). Time-series analysis of behavioural data: An update. *Behaviour Change, 4,* 40–45.
Sharpley, C. F., & Alavosius, M. P. (1988). Autocorrelation in behavioral data: An alternative perspective. *Behavioral Assessment, 10,* 243–251.
Skinner, B. F. (1938). *The behavior of organisms.* New York: Appleton-Century-Crofts.
Strupp, H. H. (1996). The tripartite model of the *Consumer Reports* study. *American Psychologist, 51,* 1017–1024.
Suen, H. K. (1987). On the epistemology of autocorrelation in applied behavior analysis. *Behavioral Assessment, 9,* 113–124.
VandenBos, G. R. (1996). Outcome assessment of psychotherapy. *American Psychologist, 51,* 1005–1006.
Watson, J. B. (1925). *Behaviorism.* New York: Norton.
Westen, D., & Bradley, R. (2005). Empirically supported complexity. *Current Directions in Psychological Science, 14,* 266–271.
Westen, D., & Morrison, K. (2001). A multidimensional meta-analysis of treatments for depression, panic, and generalized anxiety disorder: An empirical examination of the status of empirically supported therapies. *Journal of Consulting and Clinical Psychology, 69,* 875–899.
Westen, D., Novotny, C. M., & Thompson-Brenner, H. K. (2004). The empirical status of empirically supported psychotherapies: Assumptions, findings, and reporting in controlled clinical trials. *Psychological Bulletin, 130,* 631–663.
Wilcox, R. R. (2001). *Fundamentals of modern statistical methods: Substantially improving power and accuracy.* New York: Springer-Verlag.

(*Appendix follows*)

Appendix
Simulation Modeling Analysis for Brief Time-Series Data Streams

Numerous statistical approaches have been proposed to handle short streams of time-series data, but they fall short of providing users with reasonable power and good Type I error control in the face of autocorrelation (for a review, see Robey et al., 1999). One of these is ITSACORR (interrupted time-series analysis procedure), which attempts to analyze short time-series data streams (Crosbie, 1993) but is somewhat overly stringent with respect to Type I error control and has unacceptable power with shorter data streams. Table A1 compares Type I error rates and power across data streams of varying lengths and autocorrelation for simulation modeling analysis (SMA) and ITSACORR.

Simulation Modeling Explained With an Example

The following example demonstrates an application of the simulation modeling approach. The hypothetical dependent variable (DV) is anxiety ratings from a single patient over time:

13.1, 15.4, 11.0, 18.0, 21.0, 18.7, 13.1, 13.2, 8.0, 9.0, 3.0, 7.0, 11.0, 9.0

The independent variable (IV; phase vector) represents the baseline (0) and the onset of the intervention (1):

0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1

Table A1
Empirical Type I Error Rates and Power of Simulation Modeling Analysis (SMA) for Time-Series and ITSACORR (Crosbie, 1993)

	N = 10 (5, 5)		*N* = 20 (10, 10)		*N* = 30 (15, 15)	
AR	ITSACORR	SMA	ITSACORR	SMA	ITSACORR	SMA
			Type I error			
0	0.02	0.00	0.01	0.06	0.02	0.06
.2	0.02	0.01	0.01	0.05	0.01	0.07
.4	0.02	0.01	0.01	0.05	0.01	0.06
.6	0.03	0.01	0.01	0.05	0.00	0.06
.8	0.03	0.01	0.02	0.05	0.01	0.05
			Power (effect = 5)			
0	0.38	0.99	0.66	1.00	0.87	1.00
.2	0.43	0.99	0.65	1.00	0.81	1.00
.4	0.51	0.99	0.77	0.99	0.84	1.00
.6	0.57	0.97	0.93	0.99	0.97	0.99
.8	0.59	0.95	0.99	0.92	1.00	0.90

Note. Both approaches offer good Type I error control in the face of autocorrelation (AR) and with very short data streams (although ideal Type I error rates should be between .025 and .075). However, SMA offers superior power to ITSACORR. Power estimates are based on the smallest effect size reported by Crosbie (1993), although SMA offers adequate power (> .80) with effects as small as 2 to 3 (not shown). ITSACORR = interrupted time-series analysis procedure.

Step 1

Calculate the correlation (Pearson *r* is used for this example) between the DV and the phase vector: $r_{\text{original}} = -.75$.

Step 2

Calculate the autocorrelation (*AR*; Lag 1 for this example) estimate of the DV: $AR = .69$.

$$AR = \frac{\frac{1}{(n-k)}\sum_{i=1}^{n-1}(x_i - \bar{x})(x_{i+1} - \bar{x})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

where n = number of data points = 14, and k = lag = 1.

Step 3

Generate a null autocorrelated distribution one data stream at a time and compare the correlation between each null autocorrelated data stream and the phase vector (r_s; where $s = 1$ to the number of simulation data streams to be generated—10,000 in this example) with r_{original}. Each data stream (14 data points each for this example) is generated using the formula below:

$$y_i = \alpha + \varepsilon_i$$

where $\varepsilon_i = \rho\varepsilon_{i-1}$ + normal random error N(0, 1); ρ = programmed autocorrelation = .69; α = programmed intercept change = 0; and i = 1 to 14.

Each resulting data stream is correlated with the original phase vector, and the absolute value of this correlation coefficient (r_s) is compared with the absolute value of r_{original}.

An example null autocorrelated data stream generated with the formula above might look like the following:

0.3312764, 0.2024504, 0.3337031, 0.9529499, −0.6504073, 0.3862382, 0.3212215, 0.3212215, −0.714993, −1.421806, −1.147819, −0.7311038, 0.9587694, 1.613709.

The correlation between this data stream and the phase vector is $r_1 = -.26$.

$|-.26| < |-.75|$: so this is counted as a "miss."

This process is repeated with a new null autocorrelated data stream:

1.333145, 1.137773, 0.1980224, 0.3734288, 0.8907856, 0.7776737, −0.229401, −0.229401, −1.103827, 0.2006088, −1.16338, −1.006609, −0.7803311, −1.867005.

The correlation between this data stream and the phase vector is $r_2 = -.79$.

$|-.79| > |-.75|$: so this is counted as a "hit."

Step 4

Step 3 is repeated a total of 10,000 times (r_1 to $r_{10,000}$) and the empirical p value is equal to "hits"/10,000. After the full procedure is completed 10,000 times, the resultant p value = .12, and we would fail to reject the null hypothesis.

Cross-Corrrelations

Cross-correlation empirical significance can be assessed in a similar manner. However, instead of using a phase vector to correlate with the random autocorrelated data streams, pairs of random streams are cross-correlated across a series of lags.

First, the range of lags that the user is interested in is determined (e.g., −5 to +5), and the correlation coefficients are calculated using the original data streams at each of the lags. Next, the simulation model is established with the same N and autocorrelation estimates. The cross-lagged correlation coefficients are calculated for the predetermined range of lags for each of the 10,000 pairs of simulation streams. Just as above, the cross-correlation for each lag is compared between the original data streams and each of the simulation stream pairs to determine the empirical p value. Last, the critical alpha for the analysis should be adjusted for multiple comparisons in order to correct for the number of lags of interest (e.g., −5 to +5 results in 11 lags, so critical alpha could be divided by 11).