Chapter 5

# EVALUATING THE EFFICACY OF COUPLE AND FAMILY THERAPY

*Brian R. W. Baucom and Alexander O. Crenshaw*

The fields of couple and family therapy have long and proud traditions of rigorously testing the efficacy of couple- and family-based interventions. Core aspects of these traditions are the use of sound methodological designs and sophisticated statistical analyses for evaluating both statistically significant and clinically significant change. These traditions provide a strong foundation for adapting well-accepted and widely used methods to the increasingly complex and interdisciplinary clinical challenges being addressed in current couple and family treatment development and evaluation. This chapter begins with a consideration of the evolving context of treatment outcome research in couple and family therapy and presents an integrative conceptual model of relational, psychological, and physical health outcomes that are common targets in current work in couple- and family-based interventions. We then turn to a discussion of current methodological, measurement, and statistical issues in couple- and family-based intervention research and provide recommendations for considering amongst alternatives. We close with recommendations for future methodological development in couple and family therapy research. For ease of communication, we focus primarily on couple therapy throughout the chapter, as very similar issues are relevant for both couple and family therapy research.

## THE EVOLVING CONTEXT OF TREATMENT OUTCOME RESEARCH IN COUPLE THERAPY

There is a wealth of evidence supporting the efficacy of couple therapy for a range of clinical outcomes such as relationship distress and depression (e.g., D. H. Baucom, Shoham, Mueser, Daiuto, & Stickle, 1998; Snyder, Castellani, & Whisman, 2006). This empirical support is largely provided by a group of randomized clinical trials (RCTs) wherein participants were randomly assigned either to a waitlist or an experimental treatment condition or to one of two treatment conditions. Stemming from this classic design, the relative efficacy of the conditions is commonly compared using classic statistical approaches, such as repeated-measures analysis of variance for testing differential efficacy in creating statistically significant change in continuous outcomes, and Kruskal Wallis tests or ordinal regression for testing differential efficacy in ordinal outcomes such as clinically significant change categories.

One key element of this classic design and approach to testing differential efficacy, as implemented in the couple and family therapy literature, is that outcome variables are most commonly examined in isolation. For example, a treatment outcome study of couple therapy may

http://dx.doi.org/10.1037/0000101-005
*APA Handbook of Contemporary Family Psychology: Vol. 3. Family Therapy and Training*, B. H. Fiese (Editor-in-Chief)

examine changes in relationship satisfaction in one analysis and changes in communication behavior in another analysis. This approach to examining different outcomes in separate analyses is valuable for conceptual, statistical, and dissemination reasons. First, it allows for a clear focus on a primary outcome variable as well as for a within-study test of the replicability of treatment effects across outcomes. Second, the simplicity of these statistical methods permits fully powered tests of efficacy in relatively small samples. Third, the familiarity of the statistical models involved eases communication of study results.

In addition to these advantages, the widespread use of this approach to evaluating efficacy is also likely to have been influenced by two additional historical factors: (a) the funding priorities of government agencies and (b) researchers' awareness of and access to statistical methods for modeling multivariate change in simultaneous outcomes. The 1980s and early 1990s were a period of rapid growth for couple therapy research focused on treating and preventing relationship distress. The explosion in couple therapy research during this period was largely driven by marital distress being an identified priority area for government funding agencies, and relationship distress was the primary outcome variable in many couple therapy studies.

Likewise, analysis of individual treatment outcomes using univariate statistical models was consistent with the relative nascence of statistical methods for evaluating multiple outcomes simultaneously. Statistical methods for examining multivariate change over time that are much more commonplace now, such as multilevel modeling (MLM), structural equation modeling (SEM) and generalized estimating equations (GEE), were not widely introduced into the broader field of relationship science, much less the specific field of couple therapy research, until many trials of couple therapy had been conducted (see Volume 1, Chapter 17, this handbook).

Although this approach continues to figure prominently in the field of couple therapy research, the scope of couple-based interventions broadening to include more and varied forms of psychopathology and physical illness, combined with advances in multivariate statistical models, creates a tremendous opportunity for couple therapy researchers to accelerate the pace of treatment development by considering alternatives to the RCT, single outcome efficacy evaluation design. One of the primary reasons why consideration of alternative designs and analytic methods is so important is that the conditions being targeted in much current couple therapy research are frequently comorbid and share overlapping risk factors (e.g., Smith, Baron, & Grove, 2014). Examples of such conditions include, but are not limited to, mood and anxiety disorders (e.g., Clarke & Currie, 2009), substance-use disorders (e.g., Whisman, 2007), eating disorders (e.g., Bulik, Baucom, Kirby, & Pisetsky, 2011), chronic pain (e.g., Cano, Gillis, Heinz, Geisser, & Foran, 2004), cardiovascular disease (e.g., Smith et al., 2014), and metabolic syndrome (e.g., Whisman, Uebelacker, & Settles, 2010).

There are numerous conceptual models of these comorbidities and their shared risk and protective factors (e.g., Burman & Margolin, 1992; Robles, Slatcher, Trombello, & McGinn, 2014; Smith et al., 2014); however, one aspect of these conceptual models that limits their utility for advancing the development of couple therapies is that they were not created from the perspective of a couple therapist. Figure 5.1 presents a conceptual model that integrates existing models of comorbidities and shared risk factors within a couple therapy framework.

This model suggests that there is an eliciting event or circumstance that prompts a couple to seek out couple-based treatment. Drawing on Karney and Bradbury's (1995) vulnerability–stress–adaptation model, we suggest that eliciting events and circumstances can be categorized into enduring vulnerabilities (i.e., longstanding traits and/or life experiences that increase risk for illness) and stressors (i.e., distressing events and/or circumstances that require a response from the couple). The eliciting event or circumstance is thought to provoke a response from the couple, referred to as an adaptive mechanism in the model. Adaption involves coordinated biological (e.g., stress response), behavioral (e.g., communication behavior), and cognitive (e.g., attributions) responses that are efforts to reduce distress and

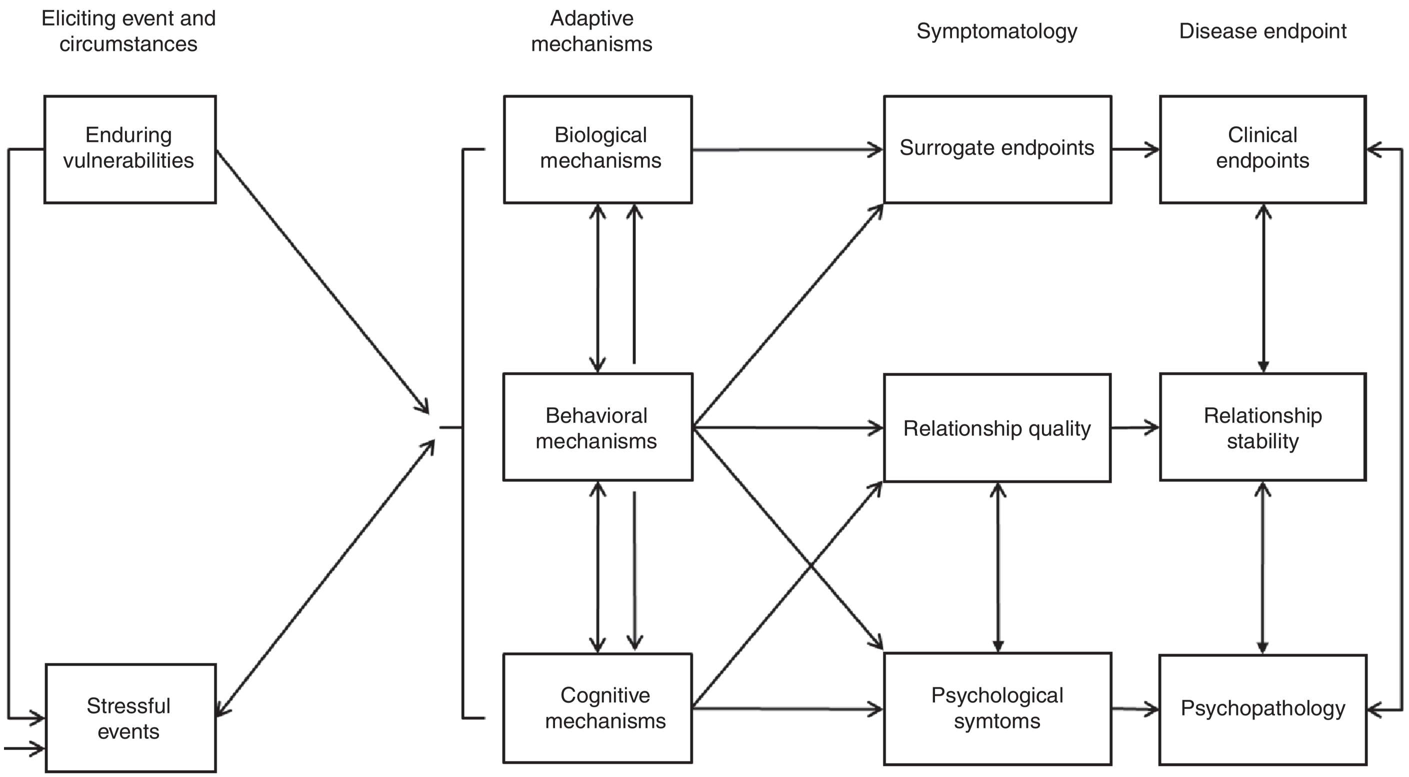

FIGURE 5.1. Heuristic model of the conditions being targeted in ongoing development of couple-based interventions. Data from Karney and Bradbury, 1995; Robles, Slatcher, Trombello, and McGinn, 2014.

to resolve the eliciting event or circumstance. Although many couples will be able to mount an effective response to a wide range of stressful events, couples that seek couple therapy are typically those who are not able to effectively cope with or to resolve the eliciting event or circumstance on their own. Ineffective responses combined with the persistence of the eliciting event or circumstance lead to increased symptoms that may manifest as physical disease progression (e.g., increased artery calcification; Robles et al., 2014; see Volume 2, Chapter 9, this handbook), increased marital distress (Karney & Bradbury, 1995), and/or increased symptoms of psychopathology (e.g., depression; Beach, Fincham, & Katz, 1998). If symptoms continue to intensify, the couple or the members of it are likely to meet criteria for a physical (e.g., congestive heart failure), relational (divorce; see Chapter 2, this volume), or psychological (e.g., depression) diagnosis.

A thorough review of the supporting empirical evidence is beyond the scope of this chapter. Interested readers are directed to Burman and Margolin (1992), Robles et al. (2014) and Smith et al. (2014) for reviews of related material. We offer this model primarily as a conceptual heuristic for contextualizing the methodological, measurement, and statistical advancements and alternatives that we describe in the remainder of the chapter below.

## METHODS FOR EVALUATING OUTCOMES OF COUPLE AND FAMILY THERAPY

Regardless of what condition is being targeted by a given couple therapy, a common series of design decisions must be made in any efficacy study. These decisions include the study design, the selection and measurement of outcome variables, and the analytic strategy for evaluating the efficacy of the experimental treatment. We review a range of options for each of these decisions below and include references for additional discussion of the issues involved.

### Study Design

There are three primary designs for evaluating the efficacy of couple therapies: single case or small "N" trials, RCTs, and open or effectiveness trials. Although RCTs are often considered the gold standard in clinical research, there are benefits and drawbacks to each type of design, depending on the aims of the study and the stage of development of the treatment. Single case or small N trials are beneficial in the early stages of treatment development and for identifying potential mechanisms, RCTs are well suited for maximizing internal validity and establishing a treatment as efficacious across groups, and effectiveness trials are ideal for determining if and to what extent treatments can be delivered effectively in real-world settings. Whereas past treatment development has followed a top-down approach in which a comprehensive treatment package is developed, evaluated for efficacy, and then subjected to dissemination efforts and dismantling designs to evaluate the components that make up the package, treatment can also follow a bottom-up design in which treatment components are tested individually at smaller scales, an approach advocated for by Christensen, Baucom, Vu, and Stanton (2005). Additionally, as the field moves beyond the "what works" model of comparing treatment packages using prepost designs and toward answering how, why, and for whom questions, an essential element to any of the designs we describe below is the use of numerous measurement occasions to examine change over time.

**Single case and small N trials.** Single case study or small N trials involve testing an intervention with one case or a small number of cases. New or experimental treatment packages or components can be tested using single case or small N designs to first determine the treatment's utility at an individual level, and modifications to treatment components or protocols can be easily made at this stage. Once a treatment is shown to be beneficial for a small number of people, treatment elements are standardized and then tested at the group level, often first using RCTs to demonstrate the treatment's generalizability to the target population and to establish more precise and stable estimates of the treatment's effect size. Following a tightly controlled RCT, the generalizability of the treatment to real world clinical settings is then ideally tested with effectiveness trials.

In addition to its utility as a means for establishing proof of concept and an opportunity to refine intervention techniques and therapist training materials before scaling up to a large N RCT, recent statistical developments have created opportunities for using small N trials to test complex models of therapeutic effects across multiple outcomes. These developments mean that it is no longer the case that vast resources are necessary to test new treatments or modifications to existing treatments, yet these models are underutilized. It is not entirely clear why there are so few examples of published small N trials of couple-based therapies, but one possible reason is the perception that it is difficult to use small N trials to establish the generalizability of a treatment. From a statistical perspective, generalizability refers to estimating between-group effects (i.e., the extent to which an association between a predictor and an outcome is consistent or variable for multiple groups). The issue with estimating between-group effects in small N trials is that 20 or more groups are generally recommended for generating stable estimates of between-group effects using modern techniques like MLM (e.g., Maas & Hox, 2005), and 30 or more groups are recommended for using Bayesian structural equation modeling (BSEM; Lee & Song, 2004).

Recent developments in time series analysis offer one potential solution to this sample size barrier. Time series analyses are regression based models that are used to analyze data from one group—a couple receiving couple therapy in this case—wherein the availability of multiple repeated measurements is leveraged to generate stable estimates of within-group change over time. One or more than one outcome variables can be analyzed in these models (e.g., time series panel analysis [TSPA]; Ramseyer, Kupper, Caspar, Znoj, & Tschacher, 2014) and multiple groups can be analyzed using an extension of classic time series analyses called pooled time series analysis (PTSA; Hoeppner, Goodwin, Velicer, & Heltshe, 2007). PTSA allows for consistency in within-group effects (i.e., changes in outcome variables over time within a couple) to be examined and tested for multiple couples within the sample model while requiring substantially fewer couples to estimate between-group effects than other alternatives like MLM or BSEM. For example, TSPA could be used to examine how a couple-based intervention created changes in relationship distress and depression for each of four couples undergoing treatment, and the consistency of the changes observed in relationship distress and depression across the four couples could be tested using PTSA. The combination of these two time series-based techniques creates a powerful opportunity for couple therapy researchers to increase the generalizability of small N trials that target one or more outcomes.

**RCTs.** RCTs utilize random assignment to directly compare two or more treatment conditions in order to establish a treatment's generalizability to the target population and understand important between-person variables. Usually, an experimental treatment is compared with a waitlist control condition, treatment as usual, or a bona fide treatment that has previously been shown to be efficacious. Evidence for an experimental intervention's efficacy can be provided by establishing superiority to treatment as usual or a waitlist control, or by demonstrating equivalency with an established treatment for the population and outcome of interest. Researchers may also consider the use of established outcome norms for common control conditions (e.g., D. H. Baucom, Hahlweg, & Kuschel, 2003) in place of an actual control condition.

To minimize the influence of confounding factors such as sample demographics, comorbidities, and treatment adherence, RCTs historically have exerted high levels of control over study parameters, such as inclusion and exclusion criteria and therapist adherence and competence in delivering treatments. One benefit of doing so is greater ability to test causal hypotheses about a treatment's efficacy by limiting alternative explanations for outcome differences (e.g., by randomly distributing individual characteristics via randomization, excluding factors such as psychosis that may impede ability to participate in and benefit from treatment, or ensuring treatments are delivered as intended). The cost of high levels of control is often that RCTs introduce characteristics of treatments that typically do not occur in nonstudy settings (e.g., Kazdin, 2008). For example, RCTs typically have a fixed number of sessions, have more restrictive inclusion and exclusion criteria than community treatment settings, and have greater standardization of treatment through use of manuals and supervision of clinicians. Participants also typically enter treatment through different avenues than couples entering treatment in the community, and they are more closely monitored than in community settings. This increased control raises questions about the extent to which results from RCTs generalize to real-world settings. For example, the few trials that have examined couple therapy in real-world settings typically have smaller effect sizes than RCTs (see Doss et al., 2012).

Multiphase optimization strategy (MOST) and sequential, multiple assessment, randomized trial (SMART) designs are two recent methodological advancements aimed at addressing concerns about the generalizability of classic RCT designs to the complexities of real-world settings (e.g., Fava, Tomba, & Tossani, 2013). Whereas RCT designs typically set a predetermined number of sessions and therapists are constrained to a single treatment package, therapists in real-world settings constantly face decision points in therapy that undoubtedly impact outcomes (Collins, Nahum-Shani, & Almirall, 2014). When faced with early nonresponse to treatment, do you stay the course or change direction? What order of interventions is optimal? Understanding these questions is key to optimizing treatment delivery in terms of time, cost, and general resource allocation, yet traditional RCTs are unable to do so. The MOST framework and SMART design use randomization at various time points—in contrast to randomization used only at the start of the study, as in typical RCTs—to determine optimal decisions at points over the course of treatment. For example, a SMART modification to Christensen et al.'s (2004) RCT comparing traditional behavioral couple therapy (TBCT; Jacobson & Margolin, 1979) with integrative behavioral couple therapy (IBCT; Jacobson & Christensen, 1998) could involve randomly assigning half the participants to each condition prior to treatment onset, then halfway through treatment randomly assigning half the nonresponders in each condition to the alternative condition and half to continued treatment in that condition. This modification would allow a test of whether nonresponders in one treatment would respond better to the other treatment.

As couple-based interventions continue to expand in scope and to be used to treat a broader range of psychological and physical illnesses, a key issue will be determination of how to integrate these couple-based interventions with existing treatment options to maximize the efficiency and flexibility of treatment regimens. Such treatments will likely involve multidisciplinary collaboration with several treatment providers. For example, a recently developed couple-based intervention for anorexia nervosa involved a treatment team of psychiatrists, nutritionists, and couple therapists (Bulik et al., 2011). Some couple-based interventions may become a first-line treatment for a disorder; others may be one of a set of equivalently efficacious treatment options for a disorder; and still others may become adjunctive treatments that are indicated for cases that fail to improve after a course of individual therapy and/or pharmacotherapy, or that can be used to enhance treatment gains after initial treatment with some other form of intervention. MOST and SMART designs hold great promise for testing amongst these various possibilities.

**Open or effectiveness trials.** In contrast to RCTs, open trials or effectiveness trials seek to maximize the generalizability of findings to real-world settings. Effectiveness trials are typically conducted in existing clinical settings and have limited or no exclusion criteria. Effectiveness trials may or may not involve randomization to study condition. One example of a large-scale effectiveness study that does involve randomization is the Veterans Administration's (VA) open trial of individual psychotherapies, comparing prolonged exposure with cognitive processing therapy for posttraumatic stress disorder (PTSD; Schnurr et al., 2015). Veterans with PTSD are not actively recruited for the study; rather, veterans who are already seeking treatment for PTSD are provided with the option to enroll in the study, and those who agree are then randomized to one of the two treatments. Often, experimental treatments are first tested using the more controlled RCT design in order to establish efficacy under optimal conditions, and then later, once evidence of efficacy has been established, the treatments are tested using effectiveness designs, so as to examine the treatment in environments that are closer to real-world treatment settings. Therefore, an effectiveness trial can be seen as a bridge between research and practice settings (Baker, McFall, & Shoham, 2008). Although they are often discussed separately, RCTs and effectiveness trials do not represent a dichotomy, but instead are part of a continuum (Baker et al., 2008); where a study falls on the continuum depends on the extent to which it prioritizes internal validity (closer to efficacy) or external validity (closer to effectiveness). However, Christensen et al. (2005) argued convincingly that it is possible to optimize both internal and external validity in clinical trials and that the two are not mutually exclusive.

Large-scale effectiveness trials have been rare in the field of couple therapy. One possible explanation for why there have been so few published effectiveness trials is that there have historically been few large-scale treatment providers who offer couple-based interventions. The VA is one exception to this general trend in that it is one of the largest providers of psychological treatment within the United States and in that it offers an expanding range of couple-based interventions. Indeed, the largest effectiveness trial of couple therapy of which we are aware was conducted in the VA (Doss et al., 2012). Additional effectiveness research is clearly one of the pressing needs for future couple therapy research.

## Selecting and Assessing Outcome Variables

In addition to selecting the study design, determining which constructs will be assessed as primary and secondary outcomes, and what methods will be used to measure those constructs are crucial decisions in efficacy research. As is true of construct measurement in general, there are numerous methods for assessing most constructs, and the choices of which particular construct is measured and the method of measuring it are influenced by numerous factors, including the theoretical model of dysfunction that is guiding the study, the study design, and the available resources. Selection of constructs and methods of measurement is also invariably influenced by a research team's areas of expertise. Self-report measures of a wide range of relationship and individual functioning constructs, and observational assessment of communication behavior and affective expression, are widely represented in couple therapy efficacy research and are methods that are likely to be familiar to many couple therapy researchers (see Chapter 3, this volume). Psychophysiological outcomes are extremely rare in couple therapy research and are likely to be less familiar to couple therapy researchers, perhaps indicating less familiarity with these methods than with self-report and observational coding. In this section, we first present a model of both the conceptual points of overlap and the unique elements of these three methods, then provide a brief overview of self-report and observational coding methods, followed by a more in-depth discussion of a conceptual model for integrating psychophysiological outcomes into future couple therapy research.

**Conceptual model of self-report, observational coding, and psychophysiological measures.** Figure 5.2 presents a Venn diagram representation of the common methods of measurement for constructs of frequent interest in couple therapy outcome research and how they relate to internal and outwardly observable processes. This conceptual model is an adaptation of Scherer's (2009) highly influential component process model of emotion (CPME). Many aspects of the CPME are relevant for constructs commonly represented in couple therapy research, with minor adaptation. In this model, a distinction is drawn between internal processes that are generally not outwardly observable and behavior that is outwardly observable. Internal processes are further divided into unconscious or automatic internal processes (e.g., physiological reactivity during a conflictual conversation with a romantic partner) and conscious or effortful internal processes (e.g., trying to understand a romantic partner's perspective during a conflictual conversation). The model also suggests that self-report measures assess the point of overlap between internal processes and observable behavior. This suggestion makes a strong assumption about

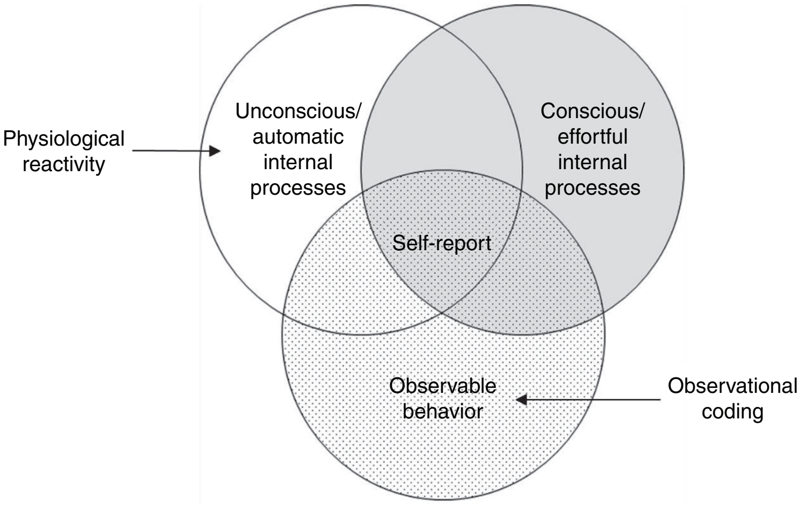
Physiological reactivity
Unconscious/ automatic internal processes
Conscious/ effortful internal processes
Self-report
Observable behavior
Observational coding

FIGURE 5.2. **Conceptual model of modalities of assessment. Data from Scherer, 2009.**

the nature of self-report measures, namely that they are representative of internal processes that the respondent is both aware and unaware of, and that some of what respondents report is to a certain extent observable to others. Another assumption about self-report as shown in this diagram is that it has unique explanatory value not captured by observational coding or physiology. The reverse is also true: Some of the observable behaviors captured by coding and some of the unconscious or automatic internal processes captured by physiology are not captured by self-report.

**Self-report measures.** Owing to the ease of delivery and the conceptual benefit of directly measuring an individual's experience, self-report measurement is perhaps the most frequently used method for measuring primary outcomes in psychological intervention research in general. Self-report measures of primary outcomes are also widely used in couple intervention research, largely because of the central role of relationship satisfaction as a primary outcome measure. Other constructs commonly measured via self-report include relationship stability (e.g., Weiss & Cerreto, 1980), communication (e.g., Crenshaw, Christensen, Baucom, Epstein, & Baucom, 2017), aggression (e.g., Straus, Hamby, Boney-McCoy, & Sugarman, 1996), process measures such as therapeutic alliance, and individual functioning variables.

Self-report measurement will likely remain a mainstay method for measuring primary outcome targets. As couple therapy research continues to expand in scope, it will be important for treatment researchers to be mindful of issues that have arisen in other treatment outcome literatures, which are likely to arise with greater frequency in couple therapy research as primary outcome targets move further afield from relationship satisfaction. One problem that is likely to have increasing relevance for couple therapy researchers is the fact that multiple primary outcome measures assessed via self-report often do not agree with one another in speaking to a treatment's efficacy, and that two treatments may appear equivalent in one domain but not in another (Achenbach, 2006; De Los Reyes, Kundey, & Wang, 2011). For example, in their widely cited study comparing dialectical behavior therapy with community treatment by experts, Linehan et al. (2006) found significant differences between treatments in suicide-related behaviors but not suicidal ideation or depression. This issue is problematic, as it suggests inconsistent evidence regarding a treatment's efficacy, and it makes comparing relative effectiveness of treatments across studies exceptionally difficult when using univariate models.

As De Los Reyes et al. (2011) noted, a common method for addressing these issues is to use a single primary outcome measure to globally assess a treatment's effectiveness along with secondary measures that are of interest but are not used in determining the efficacy of the treatment. Scholars have identified numerous problems inherent to this approach, including the following: (a) researchers often spin results when there are null findings on the primary measure, by focusing on significant results on secondary measures; (b) significant results on the primary measure are often treated as indicating global improvement in functioning, despite the measure capturing only one or a few constructs; and (c) examining multiple measures can paint a more nuanced picture of how the treatment works (e.g., De Los Reyes & Kazdin, 2008; De Los Reyes et al., 2011). To overcome these issues, De Los Reyes et al. (2011) recommended considering multiple outcome measures in tandem, a suggestion from which future couple intervention research would likely benefit.

**Observational coding.** Observational coding is a method for measuring partners' behaviors during interactions with one another (see Volume 1, Chapter 16, this handbook). It typically involves recording partners engaging in a task with one another and having trained research assistants rate the frequency and/or intensity of a set of defined behaviors during the recording. Observational coding of behavior is generally considered to be a more objective method for measuring behavior than other alternatives, such as self-reports of behavior, and is commonly used for assessing changes in communication behavior and/or affective expression produced by a course of couple therapy.

Numerous coding systems have been developed, and as researchers have pointed out (e.g., Bakeman & Gottman, 1997; Heyman, 2001), there is no one coding system that is universally best or optimal for couple therapy efficacy research. Rather, the choice of which coding system is used, and whether to create a new coding system for a given study, should be theoretically grounded. Important domains to consider are the level of analysis (i.e., is behavior to be measured in one summary score or in a score for each of several segments of the interaction?) and the need to adapt the coding system for the population being studied (Heyman, 2001).

In addition to these considerations, an additional issue that warrants consideration in future couple therapy research is related to the use of general couple communication coding systems versus disorder-specific coding systems in studies of treatment development for psychopathology or physical illness. As seen in Figure 5.1, behavioral mechanisms are thought to be a primary means by which couple adaptation is related to mental and physical health outcomes. However, it is not necessarily the case that relational, mental health, and physical health outcomes are associated with the same behaviors (i.e., the arrows from behavioral mechanisms to surrogate endpoints, relationship quality, and psychological symptoms may represent different behaviors; B. R. Baucom et al., 2007). Given the lack of empirical evidence testing this likelihood, future research would likely benefit from measuring enough general couple communication behaviors to provide a link to the large body of existing research that includes such measures, as well as measuring disorder-specific behaviors to permit sensitive tests of associations with surrogate endpoints and psychological symptoms.

**Psychophysiology.** *Psychophysiology* refers to measuring some aspect of physiological activity while participants are either at rest or performing a task. Within the broader field of relationship science, psychophysiological research generally falls into one of three broad categories: (a) individual differences in resting physiology, (b) task-related physiological reactivity, and (c) pathophysiology of disease progression. These three categories map respectively onto enduring vulnerabilities, biological adaptive mechanisms, and surrogate and clinical endpoints in Figure 5.1.

The concept of enduring vulnerabilities in psychophysiological research is both similar to and different from the application of the same concept to self-reported traits or previous life experiences. The similarity is in the conceptual notion of individual differences, namely that individuals can be rank ordered according to the value of some physiological or self-reported metric. The difference is in the interpretability of the rank ordering. Self-report measures are designed to be scaled such that if one participant has a higher score than another, that participant possesses a higher level of the trait being measured than does the other.[1] It is much more difficult, and at times impossible, to interpret individual differences in resting physiology in the same manner. At one extreme, individual differences in resting electrodermal activity (also known as galvanic skin response or skin conductance) are influenced by such a wide array of factors that they are generally understood to be uninterpretable. In contrast, individual differences in resting respiratory sinus arrhythmia (RSA) are well accepted as an index of individual differences in regulatory capacity; higher RSA indicates higher regulatory capacity (Thayer & Lane, 2000). For an accessible introduction to the use and interpretation of individual difference measures of psychophysiology, see Diamond and Otter-Henderson (2007).

In contrast to the interpretive complexity of physiological measures of enduring vulnerabilities, measurement and interpretation of physiological measures of biological adaptive mechanisms are much more widely accepted and well developed. We operationalize physiological measures of biological adaptive mechanisms as changes in physiological activity provoked by participating in a research task, such as a conversation with a partner or a standardized stress task (e.g., speeded mental arithmetic; Kirschbaum, Pirke, & Hellhammer, 1993). As shown in Figure 5.2, physiological changes are assumed to be related to automatic internal processes (e.g., a biologically coordinated stress response; Williams, Smith, Gunn, & Uchino, 2010), deliberate internal processes (e.g., cognitive emotion regulation strategies; Gross & John, 2003), and outwardly observable behavior (e.g., emotional expression; Mauss, Levenson, McCarter, Wilhelm, & Gross, 2005) and communication behavior (e.g., Brown & Smith, 1992). Because of the range of factors influencing physiological reactivity, current interpretive recommendations encourage a focus on the physiological system involved and careful interpretation of psychological meaning (Cacioppo & Tassinary, 1990). For example, much less is known about the psychological constructs involved in task-related changes in RSA compared with individual differences in resting RSA. However, changes in RSA are known to be strongly related to changes in parasympathetic activity (i.e., the “rest and digest” component of the autonomic nervous system; Thayer & Lane, 2000); thus, an increase in RSA during a couple’s conflictual interaction would be most appropriately interpreted as indicating parasympathetic augmentation. The *Handbook of Psychophysiology* (Cacioppo, Tassinary, & Berntson, 2007) is an excellent resource for readers interested in learning more about the conceptual and technical aspects of measuring and interpreting task-related changes in a wide range of physiological systems.

Finally, surrogate and clinical endpoints refer to biological indices that are associated with risk factors for or precursors of (surrogate endpoints), clinical indications of (clinical endpoints), or formal diagnosis of (clinical endpoints) physical illness. For example, stress, heightened systolic blood pressure, and obesity are all risk factors for cardiovascular disease (Kannel, Gordon, & Schwartz, 1971; Van Gaal, Mertens, & De Block, 2006). In the conceptual model advanced in this chapter, stress corresponds to the variable on the left side of Figure 5.1, blood pressure reactivity during couple conflict would represent a biological adaptive response to the stressor, obesity would be a surrogate endpoint, and a diagnosis of coronary heart disease would be a clinical endpoint (e.g., Smith & Ruiz, 2002). One practical (although oversimplified) guideline for understanding where a given physiological measure would fall in Figure 5.1 is that physiology measures obtained during most tasks performed in psychological research labs (e.g., psychosocial stress tasks) are biological adaptive mechanisms, whereas most measures found on a medical chart are surrogate (e.g., results of a blood panel) or clinical (e.g., diagnostic code) endpoints.

Psychophysiological outcomes have yet to be commonly integrated into couple therapy research;

[1] Self-report measures that perform in this manner are said to be factorially invariant, meaning that the factor structure of the scale is known to be equivalent for all groups of respondents. Very few self-report measures used in relationship science have been subjected to factorial invariance testing and are presumed, but not known, to be at least weakly factorially invariant. See South, Krueger, and Iacono (2009) for an example application of factorial invariance to the Dyadic Adjustment Scale (Spanier, 1976) and Meredith (1993) for additional discussion of factorial invariance.

however, the small amount of existing couple therapy research on these outcomes suggests great promise for incorporating them into future couple therapy research. In the two existing studies of which we are aware, participating in a relationship education program was associated with significant decreases in couples' salivary cortisol response during couple conflict at posttreatment relative to pretreatment (Ditzen, Hahlweg, Fehm-Wolfsdorf, & Baucom, 2011) and administration of intranasal oxytocin was found to be associated with decreased cortisol reactivity and an increased ratio of positive to negative communication behaviors during couple conflict (Ditzen et al., 2009). Both studies indicate that decreased hypothalamic–pituitary–adrenal axis (HPA axis) activity coincides with a beneficial response to a couple-based intervention, and suggest that couple-based interventions for relationship distress are likely to create changes in stress-related responding, which is a biological adaptive mechanism. Although it is likely that couple-based interventions could impact several paths involving physiological variables in Figure 5.1 (e.g., modifying how biologically based enduring vulnerabilities, such as low regulatory capacity, are associated with a partner's response to a stressful event), Ditzen's work, combined with the large base of literature documenting associations among physiological reactivity, communication behavior, and cognition (e.g., Robles et al., 2014), suggests that biological adaptive mechanisms are particularly likely to be impacted by a course of couple therapy.

## STATISTICAL METHODS FOR EVALUATING EFFICACY OF COUPLE AND FAMILY THERAPY

A final methodological decision involved in evaluating the efficacy of a couple-based therapy is the choice of a statistical method for estimating the magnitude and significance of the change created by the treatment. There are two main categories of statistical models for estimating the magnitude of change: models that estimate the amount of change created in terms of the scale of the outcome variable, and models that calculate clinically significant change categories. Clinically significant change categories are an ordered set of categories defined using a combination of the amount of reliable change (i.e., raw change from pretreatment to posttreatment, adjusted for the internal consistency or test–retest reliability of the outcome variable) and absolute score relative to clinical norms (i.e., above vs. below the threshold for clinically significant distress on the outcome variable). We review the numerous options for estimating both forms of change below.

### Models for Assessing Statistically Significant Change Over Time

There are many ways to estimate the statistical significance of change created by a couple-based therapy over time. Early couple therapy research frequently used repeated measures analysis of variance (ANOVA) or variants of ANOVA to examine one outcome variable at a time. These models are well suited to single primary outcome RCT designs, but they are subject to limitations that make them less well suited for future research. ANOVA models do not allow for missing data; they assume that measurement occasions are equally spaced and that measurements are collected at the equivalent time interval for all participants. These qualities of ANOVA models reduce their utility for more complex study/nesting designs (e.g., multiple membership models; Browne, Goldstein, & Rasbash, 2001) and more complex outcome models (e.g., multiple simultaneous outcomes). Multiple membership models are recommended for complex forms of nesting, such as those that are likely to occur during a SMART trial wherein some couples might receive therapy from one therapist the entire time (e.g., those who benefit from the initial treatment), whereas others may receive treatment from more than one therapist (e.g., those who do not show improvement from the initial treatment and are randomly assigned to another condition at some point during the study). Such models can be estimated using MLM. MLM has become increasingly popular for estimating the magnitude of statistical change in treatment outcome research in general because of its flexibility and applicability to a wide range of study designs, and MLM is a mainstay in couple therapy outcome research at present.

Thanks to their use of maximum likelihood estimation and empirical Bayesian estimates, multilevel models are flexible, allowing for (a) different numbers of participants in each condition, (b) missing data for measurement occasions without needing to drop entire cases or to impute missing values, (c) flexibility in the timing between measurement occasions within and between participants, and (d) greater flexibility in examining multiple outcome variables at once. MLM's flexibility in handling missing data also makes MLM recommended for intent-to-treat analyses (ITT), which are the gold standard in intervention research (e.g., Little et al., 2012). ITT refers to analyzing all participants who were randomized to a treatment condition regardless of whether they completed treatment or not. Finally, MLM allows for estimating nonlinear trajectories of change, which is advantageous for modeling change over qualitatively different phases of treatment (e.g., active treatment versus follow-up; for illustrative examples, see Baucom et al., 2015; Christensen et al., 2004).

There are numerous ways that MLM can be used to estimate statistically significant change in outcomes variables; the two most common methods of which are repeated measures ANOVA models and growth curve models. The MLM implementation of repeated measures ANOVA models is very similar to standard repeated measures ANOVA models in that both test change in mean levels of outcomes variables from one point to the next. The primary difference between the two methods is in how the magnitude of change is estimated. Standard repeated measures ANOVA models are estimated using the least squares methods and MLM repeated measures ANOVA models are estimated using maximum likelihood methods.

Growth curve models are like repeated measures ANOVA models in that both approaches model change in outcome measures over time. The primary difference between growth curve models and ANOVA models is that growth curve models describe change as a constant process that represents smooth increase or decline in mean levels of the outcome variable over three or more measurement occasions, whereas repeated measures ANOVA models are comparing mean level changes between two contiguous time points. In other words, repeated measures ANOVA models can be thought of as a specific parameterization of growth models that use categorical predictors to characterize change between two time points. It creates a false dichotomy to say that growth models are better than repeated measures ANOVA models or vice versa, because a repeated measures ANOVA model is one form of a growth model.

The generalized form of the basic growth curve model regresses the outcome variable onto time as represented by the following series of equations:

*Level 1 (measurement occasion):*

$$\text{Satisfaction}_{ijk} = \pi_{0jk} + \pi_{1jk} * \text{Time} + e_{ijk}$$

$$\textit{Level 2 (partner):}\quad \pi_{0jk} = \beta_{00k} + r_{0jk;}\pi_{1jk} = \beta_{10k} + r_{1jk}$$

$$\textit{Level 3 (couple):}\quad \beta_{00k} = \gamma_{000} + \mu_{00k;}\ \beta_{10k} = \gamma_{100} + \mu_{10k},$$

where *i* represents time points, *j* represents partners within a couple, and *k* represents couples. This basic model can be extended to incorporate additional predictors of the rate, or magnitude, of change over time. Predictors that change over time and are measured at the same time that the outcome variable is measured (e.g., depressive symptoms) are entered at Level 1; static, individual-level variables that are measured once during the study (e.g., personality) are entered at Level 2; and couple-level variables that are measured once during the study (e.g., family income) are entered at Level 3. The structure of MLM allows higher level variables to predict either the intercept of the lower level—constituting a main effect for that higher level variable—or any lower level slope, constituting an interaction (also known as a cross-level interaction). If a researcher wants to examine whether change over time occurs nonlinearly (e.g., quadratic), he or she can add Time$^2$ as a predictor at level 1.[2] To examine a main effect of treatment condition at the start of the study, it is entered at the couple level (Level 3) on the intercept (line 4, $\beta_{00k} = \gamma_{000} + \gamma_{001} * \text{Treatment condition} + \mu_{00k}$). To examine the impact of a personality characteristic

[2] The number of equations will increase as predictors are added at lower levels.

on outcome at the start of the study (e.g., neuroticism), it is entered at Level 2 on the intercept (line 2, $\pi_{0jk} = \beta_{00k} + \beta_{01k} * \text{Neuroticism} + r_{0jk}$). To examine the impact of a time-varying variable (e.g., depressive symptoms) on the outcome, it is entered as a predictor at Level 1 (after linear detrending; see Curran & Bauer, 2011; line 1, $\text{Satisfaction}_{ijk} = \pi_{0jk} + \pi_{1jk} * \text{Time} + \pi_{2jk} * \text{Depressive symptoms} + e_{ijk}$). Individual- or couple-level variables that are hypothesized to be related to the trajectory of change over time (e.g., does neuroticism predict slower improvement over time?) are entered as predictors of time on lines 3 ($\pi_{1jk} = \beta_{10k} + \beta_{11k} * \text{Neuroticism} + r_{1jk}$) or 5 ($\beta_{10k} = \gamma_{100} + \gamma_{110} * \text{Couple variable} + \mu_{10k}$), respectively. Growth models in MLM are incredibly flexible in this way, allowing any number of hypotheses to be tested.

For example, in examining the effect of behavioral couple therapy on vocally encoded emotional arousal, Baucom et al. (2015) found that both TBCT and IBCT reduce emotional arousal similarly by the end of treatment, but also found that couples who had received IBCT, an intervention focused on reducing reactivity by increasing acceptance, reached their peak arousal during conflict discussions earlier in the discussion and then subsequently decreased in arousal. In contrast, couples receiving TBCT increased in arousal linearly over the course of a conflict discussion after treatment. ANOVA-based models can test mean differences in relatively equivalent ways, but they are unable to test hypothesized differences in trajectories. The findings in Baucom et al. (2015) point to possible mechanisms of change in treatment and even provide a test of the treatment's theorized mechanism, highlighting the value of growth curve and other MLM-based models.

There are numerous references that provide additional instruction in constructing and estimating growth models in a MLM framework. Singer and Willett (2003) presented an excellent introduction to basic and advanced applications of MLM growth modeling. References are also available for specifying MLM growth models in different software packages, including HLM (Raudenbush & Bryk, 2002), R (Fox, 2015), SPSS (Heck, Thomas, & Tabata, 2013), SAS (Albright & Marinova, 2010), and Stata (Rabe-Hesketh & Skrondal, 2008).

The primary outcome, or univariate, MLM method can be extended for the purpose of considering several outcomes simultaneously, much like multivariate analysis of variance can extend ANOVAs to multiple outcomes, to test questions such as whether outcomes have different rates of change or whether different outcomes change in tandem or independently (Baldwin, Imel, Braithwaite, & Atkins, 2014). This method has tremendous promise but has, to date, not been used in much treatment outcome research. In their review of randomized trials published in the *Journal of Consulting and Clinical Psychology* during a 3-year span, Baldwin et al. (2014) found that only one of 60 randomized trials investigating multiple outcomes utilized multivariate methods to test for differential treatment effects on those outcomes. Baldwin et al. (2014) provided an excellent and accessible tutorial on implementing these models, which can be used to estimate using most MLM statistical packages.

## Methods of Estimating Effect Sizes in MLM

One complication of using MLM to quantify the magnitude of change over time is that it is commonly desirable to present such estimates in effect size metric, but there currently is no agreed-upon method for estimating effect sizes in MLM. Methods for estimating effect sizes in MLM is an area of ongoing research, and currently available methods are best considered to be reasonable approximations of effect sizes. Current methods generate these estimates for the MLM equivalent of total variance explained (the equivalent of the familiar $R^2$ statistic reported for ordinary least squares (OLS) repeated-measures ANOVA) and for the magnitude of individual predictors in a standardized metric (the equivalent of a standardized beta coefficient for continuous predictors and Cohen's *d* for categorical predictors in OLS regression or ANOVA models). The MLM equivalent of total variance explained is called "pseudo-$R^2$" and is estimated by correlating observed values of the outcome variables ($Y$) with model-based estimates of the outcome variable ($\hat{Y}$). The MLM equivalent of the standardized beta coefficient is estimated by multiplying the magnitude of the regression parameter variable produced by the

MLM by the ratio of the standard deviation of the predictor divided by the standard deviation of the outcome variable. Likewise, the MLM equivalent of Cohen's *d* is estimated by dividing the magnitude of the regression parameter produced by the MLM by the standard deviation of the outcome variable (Peugh, 2009). We strongly encourage caution in reporting and interpreting effect sizes based on MLM results.

### Methods for Determining Clinically Significant Change

The statistical significance of change created by a couple therapy tests whether a group-level change in the outcome variable is greater than chance, but it does not convey information about whether those differences are meaningful to the individuals undergoing treatment. The latter is referred to as clinical significance, and it is now a staple of clinical intervention research and used to supplement statistical significance. Additionally, clinical significance may point to a treatment's efficacy in case studies or small N trials, when examining between-group differences statistically is unadvisable because of insufficient power. Two criteria involved in determining whether a given client achieved clinically significant change is that the client shows reliable improvement on measures of interest (i.e., changed by an amount greater than measurement error) and that they surpass a predefined cutoff point, such that they are statistically indistinguishable from nondistressed individuals (see Lambert & Ogles, 2009, for a thorough discussion and history of clinical significance). The original method proposed by Jacobson, Follette, and Revenstorf (1984) established four categories that participants may fall into: recovered (i.e., showed reliable improvement and passed the predetermined cutoff point), improved (i.e., showed reliable improvement but did not pass the cutoff point), unchanged (i.e., met neither criteria), or deteriorated (i.e., demonstrated reliable decline; McGlinchey et al., 2002). An advantage of including clinical significance in clinical trials is that, in addition to examining mean-level treatment differences via statistical significance, researchers can also examine whether two treatments produce different clinically meaningful outcomes.[3]

Numerous methods have been proposed for evaluating clinical significance, and though some differences emerge in categorizations in studies that have compared them, scholars have concluded that they are roughly equivalent (Atkins, Bedics, McGlinchey, & Beauchaine, 2005; Bauer, Lambert, & Nielsen, 2004; Lambert & Ogles, 2009). The cumulative recommendation has been to use the Jacobson and Truax (1991) method for two-wave trials (e.g., pretreatment and posttreatment) based on its popularity, the lack of evidence for superiority of other methods, ease of computation, and availability of cutoff estimates for several instruments (Bauer et al., 2004). However, the HLM approach is the only one we are aware of that can be applied to data using more than two time points (e.g., growth curves; Speer & Greenbaum, 1995). Information on how to compute clinical significance using these and other methods can be found in the appendices of Bauer et al. (2004) and Atkins et al. (2005).

## FUTURE DIRECTIONS AND CONCLUSION

The design, methodological, and statistical advancements reviewed in this chapter create a range of new options for future research evaluating the efficacy of couple- and family-based treatments. These possibilities include examining a wider range and number of primary outcome variables for conditions with known physical, relational, and/or psychological comorbidities; greater flexibility in treatment delivery without loss of experimental control; and methods for estimating between-subject effects in small sample sizes. Each future study will clearly need to evaluate the potential benefits and costs of incorporating any of these possibilities. Perhaps the single most powerful modification that could be made in future efficacy research would be to incorporate more frequent assessment of the primary

[3] Clinical significance is not without its limitations. As noted by Lambert and Ogles (2009), limitations include (a) unavailability of normative data for the outcome variable, (b) inability to categorize people entering treatment already in the nondistressed range or those with chronic conditions, (c) arbitrariness of cutoffs for "normal" functioning, and (d) limited data on whether clinical significance categories are related to actual differences in functioning.

outcome variable. This study design element could make both time series and growth curve analysis possible, allow for ongoing evaluation of treatment gains (which would be needed for SMART designs), and permit examination of the timing and course of therapeutic gains across multiple outcomes. The ability to frequently assess outcomes is dependent on the availability of brief, psychometrically sound measures (e.g., Funk & Rogge, 2007), efficient methods for observational coding (e.g., K. J. Baucom, Baucom, & Christensen, 2012; Black et al., 2013), and rapid and unobtrusive methods for collecting psychophysiological data (e.g., Butner, Behrends, & Baucom, in press). Applied work examining the efficacy of couple-based interventions would benefit tremendously from further developments in methods for acquiring these data, and advancements in these methods would be a valuable direction in future basic research in relationship science.

## References

Achenbach, T. M. (2006). As others see us clinical and research implications of cross-informant correlations for psychopathology. *Current Directions in Psychological Science*, *15*, 94–98. http://dx.doi.org/10.1111/j.0963-7214.2006.00414.x

Albright, J. J., & Marinova, D. M. (2010). *Estimating multilevel models using SPSS, Stata, SAS, and R*. Bloomington, IN: Indiana University.

Atkins, D. C., Bedics, J. D., McGlinchey, J. B., & Beauchaine, T. P. (2005). Assessing clinical significance: Does it matter which method we use? *Journal of Consulting and Clinical Psychology*, *73*, 982–989. http://dx.doi.org/10.1037/0022-006X.73.5.982

Bakeman, R., & Gottman, J. M. (1997). *Observing interaction: An introduction to sequential analysis*. Cambridge, UK: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511527685

Baker, T. B., McFall, R. M., & Shoham, V. (2008). Current status and future prospects of clinical psychology: Toward a scientifically principled approach to mental and behavioral health care. *Psychological Science in the Public Interest*, *9*, 67–103. http://dx.doi.org/10.1111/j.1539-6053.2009.01036.x

Baldwin, S. A., Imel, Z. E., Braithwaite, S. R., & Atkins, D. C. (2014). Analyzing multiple outcomes in clinical research using multivariate multilevel models. *Journal of Consulting and Clinical Psychology*, *82*, 920–930. http://dx.doi.org/10.1037/a0035628

Baucom, B. R., Eldridge, K., Jones, J., Sevier, M., Clements, M., Markman, H., Stanley, S., Sayers, S., Sher, T., & Christensen, A. (2007). Contributions of marital distress and depression to communication patterns in distressed couples. *Journal of Social and Clinical Psychology*, *26*, 689–707.

Baucom, B. R., Sheng, E., Christensen, A., Georgiou, P. G., Narayanan, S. S., & Atkins, D. C. (2015). Behaviorally-based couple therapies reduce emotional arousal during couple conflict. *Behaviour Research and Therapy*, *72*, 49–55. http://dx.doi.org/10.1016/j.brat.2015.06.015

Baucom, D. H., Hahlweg, K., & Kuschel, A. (2003). Are waiting-list control groups needed in future marital therapy outcome research? *Behavior Therapy*, *34*, 179–188. http://dx.doi.org/10.1016/S0005-7894(03)80012-6

Baucom, D. H., Shoham, V., Mueser, K. T., Daiuto, A. D., & Stickle, T. R. (1998). Empirically supported couple and family interventions for marital distress and adult mental health problems. *Journal of Consulting and Clinical Psychology*, *66*, 53–88. http://dx.doi.org/10.1037/0022-006X.66.1.53

Baucom, K. J., Baucom, B. R., & Christensen, A. (2012). Do the naïve know best? The predictive power of naïve ratings of couple interactions. *Psychological Assessment*, *24*, 983–994. http://dx.doi.org/10.1037/a0028680

Bauer, S., Lambert, M. J., & Nielsen, S. L. (2004). Clinical significance methods: A comparison of statistical techniques. *Journal of Personality Assessment*, *82*, 60–70. http://dx.doi.org/10.1207/s15327752jpa8201_11

Beach, S. R., Fincham, F. D., & Katz, J. (1998). Marital therapy in the treatment of depression: Toward a third generation of therapy and research. *Clinical Psychology Review*, *18*, 635–661. http://dx.doi.org/10.1016/S0272-7358(98)00023-3

Black, M., Katsamanis, N., Baucom, B. R., Lee, C., Lammert, A., Christensen, A., . . . Narayanan, S. (2013). Toward automating a human behavioral coding system for married couples' interactions using speech acoustic features. *Speech Communication*, *55*, 1–21. http://dx.doi.org/10.1016/j.specom.2011.12.003

Brown, P. C., & Smith, T. W. (1992). Social influence, marriage, and the heart: Cardiovascular consequences of interpersonal control in husbands and wives. *Health Psychology*, *11*, 88–96. http://dx.doi.org/10.1037/0278-6133.11.2.88

Browne, W. J., Goldstein, H., & Rasbash, J. (2001). Multiple membership multiple classification (MMMC). *Statistical Modelling*, *1*, 103–124. http://dx.doi.org/10.1177/1471082X0100100202

Bulik, C. M., Baucom, D. H., Kirby, J. S., & Pisetsky, E. (2011). Uniting couples (in the treatment of) anorexia nervosa (UCAN). *International Journal of Eating Disorders*, *44*, 19–28. http://dx.doi.org/10.1002/eat.20790

Burman, B., & Margolin, G. (1992). Analysis of the association between marital relationships and health problems: An interactional perspective. *Psychological Bulletin*, *112*, 39–63. http://dx.doi.org/10.1037/0033-2909.112.1.39

Butner, J., Behrends, A., & Baucom, B. R. (in press). Modeling the emergence of co-regulation in social relations. In R. Vallacher, S. J. Read, & A. Nowak (Eds.), *Computational Models in Social Psychology*. New York, NY: Psychology Press.

Cacioppo, J. T., & Tassinary, L. G. (1990). Inferring psychological significance from physiological signals. *American Psychologist*, *45*, 16–28. http://dx.doi.org/10.1037/0003-066X.45.1.16

Cacioppo, J. T., Tassinary, L. G., & Berntson, G. (Eds.). (2007). *Handbook of psychophysiology*. Cambridge, England: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511546396

Cano, A., Gillis, M., Heinz, W., Geisser, M., & Foran, H. (2004). Marital functioning, chronic pain, and psychological distress. *Pain*, *107*, 99–106. http://dx.doi.org/10.1016/j.pain.2003.10.003

Christensen, A., Atkins, D. C., Berns, S., Wheeler, J., Baucom, D. H., & Simpson, L. E. (2004). Traditional versus integrative behavioral couple therapy for significantly and chronically distressed married couples. *Journal of Consulting and Clinical Psychology*, *72*, 176–191. http://dx.doi.org/10.1037/0022-006X.72.2.176

Christensen, A., Baucom, D. H., Vu, C. T. A., & Stanton, S. (2005). Methodologically sound, cost-effective research on the outcome of couple therapy. *Journal of Family Psychology*, *19*, 6–17. http://dx.doi.org/10.1037/0893-3200.19.1.6

Clarke, D. M., & Currie, K. C. (2009). Depression, anxiety and their relationship with chronic diseases: A review of the epidemiology, risk and treatment evidence. *The Medical Journal of Australia*, *190*(7, Suppl.), S54–S60.

Collins, L. M., Nahum-Shani, I., & Almirall, D. (2014). Optimization of behavioral dynamic treatment regimens based on the sequential, multiple assignment, randomized trial (SMART). *Clinical Trials*, *11*, 426–434. http://dx.doi.org/10.1177/1740774514536795

Crenshaw, A. O., Christensen, A., Baucom, D. H., Epstein, N. B., & Baucom, B. R. W. (2017). Revised scoring and improved reliability for the Communication Patterns Questionnaire. *Psychological Assessment*, *29*, 913–925. http://dx.doi.org/10.1037/pas0000385

Curran, P. J., & Bauer, D. J. (2011). The disaggregation of within-person and between-person effects in longitudinal models of change. *Annual Review of Psychology*, 62, 583–619. http://dx.doi.org/10.1146/annurev.psych.093008.100356

De Los Reyes, A., & Kazdin, A. E. (2008). When the evidence says, “Yes, no, and maybe so”: Attending to and interpreting inconsistent findings among evidence-based interventions. *Current Directions in Psychological Science*, *17*, 47–51. http://dx.doi.org/10.1111/j.1467-8721.2008.00546.x

De Los Reyes, A., Kundey, S. M., & Wang, M. (2011). The end of the primary outcome measure: A research agenda for constructing its replacement. *Clinical Psychology Review*, *31*, 829–838. http://dx.doi.org/10.1016/j.cpr.2011.03.011

Diamond, L. M., & Otter-Henderson, K. D. (2007). Physiological measures. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 370–388). New York, NY: Guilford.

Ditzen, B., Hahlweg, K., Fehm-Wolfsdorf, G., & Baucom, D. (2011). Assisting couples to develop healthy relationships: Effects of couples relationship education on cortisol. *Psychoneuroendocrinology*, *36*, 597–607. http://dx.doi.org/10.1016/j.psyneuen.2010.07.019

Ditzen, B., Schaer, M., Gabriel, B., Bodenmann, G., Ehlert, U., & Heinrichs, M. (2009). Intranasal oxytocin increases positive communication and reduces cortisol levels during couple conflict. *Biological Psychiatry*, *65*, 728–731. http://dx.doi.org/10.1016/j.biopsych.2008.10.011

Doss, B. D., Rowe, L. S., Morrison, K. R., Libet, J., Birchler, G. R., Madsen, J. W., & McQuaid, J. R. (2012). Couple therapy for military veterans: Overall effectiveness and predictors of response. *Behavior Therapy*, *43*, 216–227. http://dx.doi.org/10.1016/j.beth.2011.06.006

Fava, G. A., Tomba, E., & Tossani, E. (2013). Innovative trends in the design of therapeutic trials in psychopharmacology and psychotherapy. *Progress in Neuro-Psychopharmacology & Biological Psychiatry*, *40*, 306–311. http://dx.doi.org/10.1016/j.pnpbp.2012.10.014

Fox, J. (2015). *Applied regression analysis and generalized linear models*. Thousand Oaks, CA: Sage.

Funk, J. L., & Rogge, R. D. (2007). Testing the ruler with item response theory: Increasing precision of measurement for relationship satisfaction with the Couples Satisfaction Index. *Journal of Family Psychology*, *21*, 572–583. http://dx.doi.org/10.1037/0893-3200.21.4.572

Gross, J. J., & John, O. P. (2003). Individual differences in two emotion regulation processes: Implications for affect, relationships, and well-being. *Journal of Personality and Social Psychology*, *85*, 348–362. http://dx.doi.org/10.1037/0022-3514.85.2.348

Heck, R. H., Tabata, L., & Thomas, S. L. (2013). *Multilevel and longitudinal modeling with IMB SPSS*. Routledge.

Heyman, R. E. (2001). Observation of couple conflicts: Clinical assessment applications, stubborn truths, and shaky foundations. *Psychological Assessment*, *13*, 5–35. http://dx.doi.org/10.1037/1040-3590.13.1.5

Hoeppner, B. B., Goodwin, M. S., Velicer, W. F., & Heltshe, J. (2007). An applied example of pooled time series analysis: Cardiovascular reactivity to stressors in children with autism. *Multivariate Behavioral Research*, *42*, 707–727. http://dx.doi.org/10.1080/00273170701755291

Jacobson, N. S., & Christensen, A. (1998). *Acceptance and change in couple therapy: A therapist's guide to transforming relationships*. New York, NY: Norton.

Jacobson, N. S., Follette, W. C., & Revenstorf, D. (1984). Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. *Behavior Therapy*, *15*, 336–352. http://dx.doi.org/10.1016/S0005-7894(84)80002-7

Jacobson, N. S., & Margolin, G. (1979). *Marital therapy: Strategies based on social learning and behavior exchange principles*. New York, NY: Brunner/Mazel.

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. http://dx.doi.org/10.1037/0022-006X.59.1.12

Kannel, W. B., Gordon, T., & Schwartz, M. J. (1971). Systolic versus diastolic blood pressure and risk of coronary heart disease. The Framingham study. *The American Journal of Cardiology*, *27*, 335–346. http://dx.doi.org/10.1016/0002-9149(71)90428-0

Karney, B. R., & Bradbury, T. N. (1995). The longitudinal course of marital quality and stability: A review of theory, method, and research. *Psychological Bulletin*, *118*, 3–34. http://dx.doi.org/10.1037/0033-2909.118.1.3

Kazdin, A. E. (2008). Evidence-based treatment and practice: New opportunities to bridge clinical research and practice, enhance the knowledge base, and improve patient care. *American Psychologist*, *63*, 146–159. http://dx.doi.org/10.1037/0003-066X.63.3.146

Kirschbaum, C., Pirke, K. M., & Hellhammer, D. H. (1993). The "Trier Social Stress Test"—A tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology*, *28*, 76–81. http://dx.doi.org/10.1159/000119004

Lambert, M. J., & Ogles, B. M. (2009). Using clinical significance in psychotherapy outcome research: The need for a common procedure and validity data. *Psychotherapy Research*, *19*, 493–501. http://dx.doi.org/10.1080/10503300902849483

Lee, S. Y., & Song, X. Y. (2004). Evaluation of the Bayesian and maximum likelihood approaches in analyzing structural equation models with small sample sizes. *Multivariate Behavioral Research*, *39*, 653–686. http://dx.doi.org/10.1207/s15327906mbr3904_4

Linehan, M. M., Comtois, K. A., Murray, A. M., Brown, M. Z., Gallop, R. J., Heard, H. L., . . . & Lindenboim, N. (2006). Two-year randomized controlled trial and follow-up of dialectical behavior therapy vs. therapy by experts for suicidal behaviors and borderline personality disorder. *Archives of General Psychiatry*, *63*, 757–766.

Little, R. J., D'Agostino, R., Cohen, M. L., Dickersin, K., Emerson, S. S., Farrar, J. T., . . . Stern, H. (2012). The prevention and treatment of missing data in clinical trials. *The New England Journal of Medicine*, *367*, 1355–1360. http://dx.doi.org/10.1056/NEJMsr1203730

Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *1*, 86–92. http://dx.doi.org/10.1027/1614-2241.1.3.86

Mauss, I. B., Levenson, R. W., McCarter, L., Wilhelm, F. H., & Gross, J. J. (2005). The tie that binds? Coherence among emotion experience, behavior, and physiology. *Emotion*, *5*, 175–190. http://dx.doi.org/10.1037/1528-3542.5.2.175

McGlinchey, J. B., Atkins, D. C., & Jacobson, N. S. (2002). Clinical significance methods: Which one to use and how useful are they? *Behavior Therapy*, *33*, 529–550. http://dx.doi.org/10.1016/S0005-7894(02)80015-6

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*, 525–543. http://dx.doi.org/10.1007/BF02294825

Peugh, J. L. (2009). A practical guide to multilevel modeling. *Journal of School Psychology*, *48*, 85–112. http://dx.doi.org/10.1016/j.jsp.2009.09.002

Rabe-Hesketh, S., & Skrondal, A. (2008). *Multilevel and longitudinal modeling using Stata*. College Station, TX: STATA Press.

Ramseyer, F., Kupper, Z., Caspar, F., Znoj, H., & Tschacher, W. (2014). Time-series panel analysis (TSPA): Multivariate modeling of temporal associations in psychotherapy process. *Journal of Consulting and Clinical Psychology*, *82*, 828–838. http://dx.doi.org/10.1037/a0037168

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Robles, T. F., Slatcher, R. B., Trombello, J. M., & McGinn, M. M. (2014). Marital quality and health: A meta-analytic review. *Psychological Bulletin*, *140*, 140–187. http://dx.doi.org/10.1037/a0031859

Scherer, K. R. (2009). The dynamic architecture of emotion: Evidence for the component process model. *Cognition and Emotion*, *23*, 1307–1351.

Schnurr, P. P., Chard, K. M., Ruzek, J. I., Chow, B. K., Shih, M. C., Resick, P. A., . . . Lu, Y. (2015). Design of VA Cooperative Study #591: CERV-PTSD, comparative effectiveness research in veterans with PTSD. *Contemporary Clinical Trials*, *41*, 75–84. http://dx.doi.org/10.1016/j.cct.2014.11.017

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford, England: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780195152968.001.0001

Smith, T. W., Baron, C. E., & Grove, J. L. (2014). Personality, emotional adjustment, and cardiovascular risk: Marriage as a mechanism. *Journal of Personality*, *82*, 502–514. http://dx.doi.org/10.1111/jopy.12074

Smith, T. W., & Ruiz, J. M. (2002). Psychosocial influences on the development and course of coronary heart disease: Current status and implications for research and practice. *Journal of Consulting and Clinical Psychology*, *70*, 548–568. http://dx.doi.org/10.1037/0022-006X.70.3.548

Snyder, D. K., Castellani, A. M., & Whisman, M. A. (2006). Current status and future directions in couple therapy. *Annual Review of Psychology*, *57*, 317–344. http://dx.doi.org/10.1146/annurev.psych.56.091103.070154

South, S. C., Krueger, R. F., & Iacono, W. G. (2009). Factorial invariance of the Dyadic Adjustment Scale across gender. *Psychological Assessment*, *21*, 622–628. http://dx.doi.org/10.1037/a0017572

Spanier, G. B. (1976). Measuring dyadic adjustment: New scales for assessing the quality of marriage and similar dyads. *Journal of Marriage and the Family*, *38*, 15–28. http://dx.doi.org/10.2307/350547

Speer, D. C., & Greenbaum, P. E. (1995). Five methods for computing significant individual client change and improvement rates: Support for an individual growth curve approach. *Journal of Consulting and Clinical Psychology*, *63*, 1044–1048. http://dx.doi.org/10.1037/0022-006X.63.6.1044

Straus, M. A., Hamby, S. L., Boney-McCoy, S., & Sugarman, D. B. (1996). The revised conflict tactics scales (CTS2) development and preliminary psychometric data. *Journal of Family Issues*, *17*, 283–316. http://dx.doi.org/10.1177/019251396017003001

Thayer, J. F., & Lane, R. D. (2000). A model of neurovisceral integration in emotion regulation and dysregulation. *Journal of Affective Disorders*, *61*, 201–216. http://dx.doi.org/10.1016/S0165-0327(00)00338-4

Van Gaal, L. F., Mertens, I. L., & De Block, C. E. (2006). Mechanisms linking obesity with cardiovascular disease. *Nature*, *444*, 875–880. http://dx.doi.org/10.1038/nature05487

Weiss, R. L., & Cerreto, M. C. (1980). The Marital Status Inventory: Development of a measure of dissolution potential. *American Journal of Family Therapy*, *8*, 80–85. http://dx.doi.org/10.1080/01926188008250358

Whisman, M. A. (2007). Marital distress and *DSM–IV* psychiatric disorders in a population-based national survey. *Journal of Abnormal Psychology*, *116*, 638–643. http://dx.doi.org/10.1037/0021-843X.116.3.638

Whisman, M. A., Uebelacker, L. A., & Settles, T. D. (2010). Marital distress and the metabolic syndrome: Linking social functioning with physical health. *Journal of Family Psychology*, *24*, 367–370. http://dx.doi.org/10.1037/a0019547

Williams, P. G., Smith, T. W., Gunn, H. E., & Uchino, B. N. (2010). Personality and stress: Individual differences in exposure, reactivity, recovery, and restoration. In R. J. Contrada & A. Baum (Eds.), *The handbook of stress science: Biology, Psychology & Health*, (pp. 231–246). New York, NY: Springer.