Conceptual and Statistical Issues in Couples Observational Research:  
Rationale and Methods for Design Decisions

Brian R.W. Baucom¹, Karena Leo¹, Colin Adamo¹, Panayiotis Georgiou², and Katherine J.W. Baucom¹

¹ Department of Psychology, University of Utah

² Department of Electrical Engineering, University of Southern California

Author note

Brian R.W. Baucom, Department of Psychology, University of Utah; Karena Leo, Department of Psychology, University of Utah; Colin Adamo, Department of Psychology, University of Utah; Panayiotis Georgiou, Department of Electrical Engineering, University of Southern California; Katherine J.W. Baucom, Department of Psychology, University of Utah.

Preparation of this manuscript was supported in part by the Office of the Assistant Secretary of Defense for Health Affairs through the Psychological Health and Traumatic Brain Injury Research Program under Award No. W81XWH-15-1-0632. The U.S. Army Medical Research Acquisition Activity, 820 Chandler Street, Fort Detrick MD 21702-5014 is the awarding and administering acquisition office. Collection of the data presented in this manuscript was supported by a F31 NRSA fellowship from NICHD (F31HD062168), a Tamar Diana Wilson Grant from the UCLA Chicano Studies Research Center, and a Randy Gerson Memorial Grant from APF awarded to Katherine J. W. Baucom. Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the Department of Defense, National Institutes of Health, or other funding agencies.

Correspondence should be addressed to Brian R.W. Baucom, University of Utah, 380 S.
Abstract

Observational behavioral coding methods are widely used for the study of relational phenomena. There are numerous guidelines for the development and implementation of these methods that include principles for creating new and adapting existing coding systems as well as principles for creating coding teams. While these principles have been successfully implemented in research on relational phenomena, the ever expanding array of phenomena being investigated with observational methods calls for a similar expansion of these principles. Specifically, guidelines are needed for decisions that arise in current areas of emphasis in couple research including observational investigation of related outcomes (e.g., relationship distress and psychological symptoms), the study of change in behavior over time, and the study of group similarities and differences in the enactment and perception of behavior. This manuscript describes conceptual and statistical considerations involved in these three areas of research and presents principle- and empirically-based rationale for design decisions related to these issues. A unifying principle underlying these guidelines is the need for careful consideration of fit between theory, research questions, selection of coding systems, and creation of coding teams. Implications of (mis)fit for the advancement of theory are discussed.

Keywords: observational coding, longitudinal research, cross-cultural research, interrater reliability
Advances in Methods for Measuring Behavior in Relationship Science

The observation and measurement of behavior has long been a cornerstone of basic and applied relationship science (e.g., Gottman & Notarius, 2000; Heyman, 2001). Behavioral measurement continues to be a widely used methodology for studying an ever increasing range of relational, psychological, and physical health phenomena (e.g., perceived entitlement to support in individuals with chronic pain, Cano, Leong, Heller, & Lutz, 2009; person-centered care performed by spouses of individuals with dementia, Ellis-Gray, Riley, & Oyebode, 2014). As the field grows to incorporate these diverse aims, increasingly sophisticated questions are being asked using behavioral methods. These new questions create design decisions for researchers that are not addressed by existing guidelines. Thus, new guidelines are needed to provide a framework for considering the conceptual, measurement, and statistical issues involved and a means of evaluating the conceptual fit between research question, hypothesis, observational coding system, and observational coding team. The aims of this manuscript are to describe methodological issues involved in three major areas of ongoing observational research and to propose a set of guidelines for making design decisions related to these issues. We begin with an overview of current guidelines for couple observational research to provide a context for conceptual, measurement, and statistical issues that arise in three major categories of ongoing observational research. We then introduce principles and statistical methods for deciding between alternatives for addressing these issues. We close with consideration of the implications of (mis)fit for the advancement of basic and applied couples research.

Overview of current guidelines for observational methods

Current guidelines for conducting observational research on couples offer a wealth of well-established principles for designing observational coding systems and ensuring fit between
relational theory and the behaviors being measured\(^1\). More than 20 years ago, Roger Bakeman and John Gottman (1997) remarked:

We sometimes hear people ask: Do you have a coding scheme I can borrow? This seems to us a little like wearing someone else’s underwear. Developing a coding scheme is very much a theoretical act, one that should begin in the privacy of one’s own study, and the coding scheme itself represents an hypothesis, even if it is rarely treated as such. After all, it embodies the behaviors and distinctions that the investigator thinks important for exploring the problem at hand. (p. 15).

Their point is as true today as it was when these words were first printed. An observational coding system should be based on a theoretical premise and specifically intended to test a hypothesis derived from that theory (Heyman, 2001; Margolin et al., 1998). Viewing behavioral quantification methods from this perspective suggests that these methods are most appropriately viewed as tools for theoretical refinement rather than ends unto themselves. And, by extension, decisions about what behaviors to code and how to code those behaviors should be made in such a way as to maximize a study’s ability to test theoretical premises.

A great deal of observational research has been conducted in precisely this manner. Much early observational work was an outgrowth of behaviorally-based couple therapies developed in the late 1970’s and 1980’s (e.g., Jacobson & Margolin, 1979). These therapies share some core theoretical assumptions about the etiology of relationship distress including, but not limited to, social exchange theory (Thibaut & Kelley, 1978)\(^2\). Social exchange theory suggests that

\(^1\) Interested readers are encouraged to consult Bakeman and Gottman (1997), Heyman (2001), and Margolin and colleagues (1998) for discussion of methodological recommendations not addressed in this manuscript.


© 2017, American Psychological Association. This paper is not the copy of record and may not exactly replicate the final, authoritative version of the article. Please do not copy or cite without authors permission. The final article will be available, upon publication, via its DOI: 10.1037/fam0000370
relationship distress is determined in part by the ratio of positive to negative reinforcers in the relationship. Consistent with this theoretical assumption, early observational coding systems (e.g., Hops, Wills, Patterson, & Weiss, 1972) were designed to quantify positive and negative communication behaviors exhibited during conflict. As etiological models of relationship distress expanded to include cognitive (e.g., attributions; Heyman & Vivian, 2000) and affective elements (e.g., Stover, Guerney, Ginsberg, & Schlein, 1977), observational coding systems were revised (e.g., Heyman, Weiss, & Eddy, 1995) or created (e.g., Gottman, McCoy, Coan, & Collier, 1996) to include measures of these domains.

A recent meta-analysis of the observational coding literature on romantic relationships (Woodin, 2011) concluded that most observational coding systems distinguish positive and negative behaviors. Likewise, narrative reviews of this literature agree that happy couples can be distinguished from distressed couples on the basis of these coding systems (e.g., B. Baucom & Eldridge, 2013). Numerous negative and positive behaviors are significantly associated with relationship satisfaction. The most consistent and well replicated of these findings is that distressed couples exhibit significantly stronger patterns of negative reciprocity and higher levels of demand/withdraw behavior³ than do happy couples (Heyman, 2001). In sum, in observational research focused on establishing behavioral correlates of relationship distress, the field has benefitted from the consistency with which researchers have mapped behaviors codified in coding systems onto theoretical constructs in etiological models of relationship distress.

A similar benefit has been realized in applied research using these same observational

---

³ Negative reciprocity refers to an increased likelihood that a partner responds to a negative behavior from the other with a negative behavior of his/her own. Demand/withdraw behavior refers to a cycle of behavior where one partner, the demander, nags, criticizes, or blames while attempting to create change and the other partner, the withdrawer, avoids discussion, changes the subject, or ends the conversation in an attempt to maintain the status quo.
coding systems. The conceptual fit between observational coding systems and etiological models of relationship distress also made these systems well-suited for evaluating the efficacy of cognitive behavioral couple therapies (CBCTs). Changes in observed communication behaviors that emerge over a course of a CBCT have often been examined either as a secondary outcome or as a potential mediator of improvements in relationship distress (e.g., Christensen, Baucom, Vu, & Stanton, 2005; Snyder & Wills, 1989). Meta-analytic (e.g., Shadish & Baldwin, 2005) and narrative reviews (e.g., Snyder, Castellani, & Whisman, 2006) agree that CBCTs significantly decrease negative behavior and significantly increase positive behavior. Less consistent evidence exists for the association between changes in observed communication behavior and changes in relationship satisfaction. It appears that while there is substantial evidence that behaviorally-based couple therapies increase relationship satisfaction, increase positive behaviors, and decrease negative behaviors, there is not sufficient evidence to conclude that behavioral changes are related to changes in relationship distress (K. Baucom, B. Baucom, & Christensen, 2015).

Other guidelines for creating and using observational coding systems have been less consistently and successfully implemented to date; many of these guidelines are statistical in nature. These guidelines include ensuring adequate interrater reliability of codes, testing and reporting internal reliability at the level of analysis at which coding data are analyzed, testing the construct validity of codes/composites of codes, and considering potential sources of subjective bias when recruiting coders and creating coding teams (e.g., Heyman, 2001; Margolin et al., 1998). Direct replication of results has been rare; it has been challenging for the field to reconcile conflicting findings; and the pace of theoretical refinement has been slowed as a result (Heyman, 2001). There are likely numerous reasons why these issues continue to plague observational research on couples. We suspect that one driving factor is the establishment of unspecified but
generally accepted norms for conducting and evaluating observational research that are used to
guide decisions in place of statistical and/or methodological principles. We do not mean to imply
there is no place for the wisdom of experience in decision-making about research design or that it
is problematic in and of itself. Rather, we think the field would benefit from a renewed emphasis
on incorporating statistical and methodological principles with an appreciation for novel
conceptual issues when designing future observational coding research.

**Three major design issues in current observational research**

There are several reasons why an appreciation of new conceptual issues and a renewed
emphasis on statistical and methodological issues involved in current observational research
would likely be beneficial. First, observational methods are being used to test new and complex
questions where there is less accumulated experiential knowledge. Second, there are recently
developed methods for measuring behavior that offer new options to researchers, and guidelines
are needed for selecting among these options. Third, independent of the volume of accumulated
experiential knowledge with a line of research or a particular method of quantifying behavior,
there are long-standing statistical methods that could be used to avoid limitations present in prior
research. In this section, we outline three sets of design decisions that have relevance for a wide
cross-section of current observational research. For each decision, we describe conceptual,
statistical, and methodological issues involved and offer recommendations addressing these
issues in the design of future observational research.

**Issue 1, Disentangling behavioral associations for related outcomes**

Many of the relational and individual outcomes frequently examined in relationship
science are known to be related to one another. For example, there is a well-established
correlation between higher levels of relationship distress and both higher levels of depressive
symptoms and a greater likelihood of being diagnosed with depression (e.g., Whisman, 2001). One enduring question in research on depression within the context of relationship distress is whether couple communication behavior is attributable uniquely to depression, uniquely to relationship distress, or to a combination of the two factors. Previous research in this area has focused on methodological issues related to sample composition (e.g., Biglan et al., 1985; Schmaling & Jacobson, 1990) and behavioral task design (Rehman et al., 2010). Evolving conceptual models for couple-based treatment of physical disease and psychopathology (e.g., Fischer & D. Baucom, in press) suggest that it will be equally important for future research to consider what kinds of behaviors are measured as well. Recent theoretical models suggest that couple interaction within the context of physical disease and/or psychological disorder are shaped by general relationship functioning (i.e., the level of relationship satisfaction) and disease specific mechanisms (e.g., D. Baucom, Whisman, & Paprocki, 2012).

This perspective suggests that it will not be possible to determine disease specific behavioral mechanisms without controlling for behaviors known to be associated with relationship satisfaction. We see the issue we are raising here as a multivariate extension of Heyman’s (2001) recommendation of making construct validity a primary concern when developing a new coding system. Assessing construct validity necessitates testing both convergent and discriminant validity of an instrument. As applied to observational coding for related outcomes, establishing construct validity would require partialling of outcomes and behaviors in order to have confidence in tests of convergent and discriminant validity.

For example, Figure 1 is a Venn diagram style representation of the potential main effects among relationship distress, depression, general communication behaviors, and depression specific communication behaviors. Each circle represents the variance in one of these constructs.
and the overlaps between the circles represents covariance between two or more of these constructs. The numbers in the overlaps between the circles indicate the unique forms of covariance that could exist for the set of two outcomes and two types of behaviors. Suppose that a researcher wanted to develop a new coding system that measures behaviors that are uniquely linked to depression (i.e., convergent validity) and not otherwise accounted for by relationship distress or general communication behaviors (i.e., discriminant validity). There are seven separate sources of potential covariance (overlaps 1, 2, 3, 5, 7, 8, 9) that would need to be accounted for to accurately estimate the partial correlation between depression specific behaviors and depressive symptoms (overlap 4) in a test of convergent validity. Omitting either relationship distress or general communication behaviors would likely bias the partial correlation between depression specific behaviors and depressive symptoms. Depending on the sizes of the potential sources of covariance, the partial correlation could be either up- or downwardly biased.

Based on the importance of conducting accurate tests of convergent and discriminant validity, we recommend including measures of relationship distress and established behavioral correlates of relationship distress in research on disorder and illness specific behaviors. In addition to increasing the accuracy of tests of convergent and discriminant validity, including these measures would yield more precise information for developing or refining couple-based interventions for psychological disorders and physical illnesses. It will be difficult to address the

4 Researchers may also wish to investigate interactive associations between behaviors and outcomes in tests of convergent and discriminant validity similar to the direct product method of multitrait-multimethod (MTMM) evaluation of construct validity (e.g., Baggozzi, Yi, & Phillips, 1991). Inclusion of interaction terms in tests of construct validity usually focuses on testing differential strength of association within vs. between methods of assessing constructs (e.g., common methods variance). As applied to couple observational research, inclusion of interaction terms could help to establish whether a disease/disorder specific behavior relates to disease/disorder severity uniformly across levels of relationship satisfaction or not. Such information could be valuable for developing/adapting different forms of couple-based intervention (i.e., couple, disorder-specific, and partner-assisted therapy; D. Baucom et al., 2012) to target specific behavioral mechanisms for comorbid relationship distress and a disease/disorder relative to those with only a disease/disorder and no relationship distress.

© 2017, American Psychological Association. This paper is not the copy of record and may not exactly replicate the final, authoritative version of the article. Please do not copy or cite without authors permission. The final article will be available, upon publication, via its DOI: 10.1037/fam0000370
unique needs of different conditions without a method for confidently determining unique behavioral mechanisms involved in the onset and maintenance of different conditions.

The major practical hurdle to including measures of established behavioral correlates of relationship distress and measures of disorder or illness specific behaviors is the resource intensive nature of coding both types of behaviors using standard coding methods, particularly if one system is being coded primarily as a covariate. Two recent advancements in observational coding methods, naïve coding and Behavioral Signal Processing (BSP; Narayanan & Georgiou, 2013), offer alternatives for coding established correlates of relationship distress that are more efficient than standard coding methods. The efficiency of these methods increases the viability of coding both disorder/illness specific behaviors and established correlates of relationship distress.

Naïve coding

Naïve coding is a type of observational coding methodology where coders are provided with minimal training and limited instruction. Similar to standard observational coding practice (which we will refer to as trained observational coding), naïve coding involves creating a coding manual that lists the types of behavior to be rated and that provides brief guidelines for each rating scale. Naïve coding builds on the tradition of viewing coders not as mere detectors of specific behaviors but also as “cultural informants” (Bakeman & Gottman, 1997). The cultural informant perspective suggests that coders’ life experiences are an integral and essential part of recognizing the behaviors listed in a coding manual. Naïve coding leverages this experiential knowledge by asking coders to make intuitive judgements about affective expression, positive

---

5 The distinction between naïve and trained coding is similar to the distinction between intuitive and rational judgements in Kahneman and Tversky’s work on decision making under conditions of uncertainty. In naïve coding, coders rely on accumulated life experience to judge the occurrence and strength of positive or negative behaviors. This process is similar in many ways to intuitive judgments, which are judgments that are made automatically without the need for much reflection (e.g., Kahneman, 2003). In contrast, in trained coding, coders use a set of rules

© 2017, American Psychological Association. This paper is not the copy of record and may not exactly replicate the final, authoritative version of the article. Please do not copy or cite without authors permission. The final article will be available, upon publication, via its DOI: 10.1037/fam0000370
and negative behaviors and/or more abstract behavioral constructs such as overall relationship functioning or likelihood of divorce (K. Baucom, B. Baucom, & Christensen, 2012). Although both naïve and trained observational coding are used to assess similar behavioral constructs and both involve creating a coding manual to specify the behaviors to be coded, the two approaches differ in the level of specificity with which they operationalize how to recognize the occurrence and/or to determine the intensity of a behavior. Naïve coding manuals provide a minimal amount of information in an effort to maximize coders’ abilities to make intuitive judgments whereas trained coding manuals provide substantial amounts of information in an effort to ensure that coders apply a common rubric to specified behavioral cues in a consistent manner.

A small but growing group of naïve coding studies of couple interaction provide initial evidence that positive and negative affect, positive and negative communication behaviors, and abstract measures of overall relationship functioning can be reliably coded using naïve methods and that these naively coded data have strong convergent validity. Waldinger, Hauser, Schulz, Allen, and Crowell (2004) conducted a pioneering study of naïve coding of couple interactions showing that naïve raters could reliably rate emotional expressions during 10 min interactions. Subsequent work replicated the feasibility of naïve coding of emotions during couple interaction and that different categories of emotions (i.e., hard and soft negative emotions; Sanford, 2007) and larger numbers of emotions (Roberts, Leonard, Butler, Levenson, & Kanter, 2013) could be reliably coded using naïve methodology. K. Baucom et al. (2012) demonstrated that naïve coding methods can also be used to generate reliable ratings of positive and negative communication
behaviors during couple interaction. In addition to demonstrating adequate interrater reliability, naïve codes were also significantly associated with concurrent and future relationship satisfaction. Two additional studies have demonstrated acceptable to high interrater reliability for naïve coding of positive and negative communication behaviors (Crane, Testa, Schlauch, & Leonard, 2016; Luebcke et al., 2014).

Finally, naïve coding approaches have been used to code highly abstract dimensions of relationship functioning based on couple interactions. K. Baucom et al. (2012) report an interrater reliability of .8 for naïvely rated overall relationship quality. The overall relationship quality code was found to be significantly associated with concurrent relationship satisfaction, and the percentage of variance in concurrent relationship distress accounted for by overall relationship quality was not significantly different than that accounted for by a set of four, psychometrically optimized scale scores created from highly trained coding data. Ebling and Levenson (2003) used a naïve approach to rating relationship distress and likelihood of divorce. They compared the predictive utility of these ratings across groups of coders with varying levels of personal experience with marriage and/or divorce. Raters for whom personal experience with marriage and/or divorce was more salient (i.e., those who were recently divorced, in long-term marriages, or newlyweds) were significantly more accurate in predicting level of relationship distress than were raters for whom professional experience was more salient (i.e., marital researchers, therapists, pastoral counselors, graduate students); there were no significant group differences in accuracy of rating likelihood of divorce (Ebling & Levenson, 2003).

In sum, this nascent body of evidence suggests that naive coding is a sound alternative to traditional coding methods for measuring established behavioral correlates of relationship satisfaction. Existing empirical evidence suggests that naive codes of negative reciprocity,
positive reciprocity, and demand/withdraw behavior are moderately to highly correlated with trained codes for those same behaviors. Finally, naïve codes appear to have levels of convergent validity comparable to those for trained codes of the same behaviors.

**Behavioral Signal Processing (BSP)**

Behavioral Signal Processing (BSP; Narayanan & Georgiou, 2013) refers to computational tools, stemming from the fields of Signal Processing and Machine Learning, which enable the measurement, analysis, and modeling of human behavior. BSP does with computer algorithms what trained and naïve coding approaches do with research assistants (RAs). The signals that BSP processes are the sounds in audio-recordings and pictures in video-recordings of couple interactions. Just like human coders, BSP integrates multiple modalities of behavioral information (i.e., what was said, how it was said) to estimate a score for behaviors defined in a coding system. For example, in the same way that a RA would consider voice tone and words spoken when scoring criticism, BSP integrates acoustic information, such as tone of voice, with lexical information, such as the words that were used, to estimate a criticism score.

In order to generate acoustic and lexical information for estimating a coding score, BSP performs a series of processing steps that RAs perform naturally and without effort. These steps include distinguishing speech from background noise and determining who is talking when, what they are saying, and what they are conveying with their voice in addition to their words. The acoustic and lexical information produced by these processing steps is then used to estimate observational coding scores via Machine Learning (ML) algorithms. Machine Learning refers to a class of computational techniques that involve data driven discovery of patterns that maximize the accuracy with which an algorithm can perform some task. Additional details of these algorithms and the processing steps of BSP are presented in the Appendix.
Similar to naïve coding approaches, BSP methods have been used to estimate coding values for established behavioral correlates of relationship satisfaction. Georgiou and colleagues (e.g., Black et al., 2013) have produced a series of studies showing that BSP methods can be used to estimate positive and negative emotion, positive and negative behavior, acceptance, and blame scores that are highly correlated (e.g., \( r \leq .82 \)) with coding values generated by trained RAs. Additional evidence supporting the promise of BSP methods for “coding” couple interactions comes from research using BSP methods to estimate behavior during other forms of psychological meaningful, dyadic interactions (e.g., empathy during Motivational Interviewing psychotherapy sessions; Xiao, Georgiou, Imel, Atkins, & Narayanan, 2015).

BSP methods have great promise for efficiently estimating scores for established behavioral correlates of relationship satisfaction. One substantial issue in implementing BSP methods is that they require expertise in signal processing and Machine Learning. Most doctoral programs in psychology do not offer training in these subjects so successful implementation of BSP methods typically involves interdisciplinary collaboration with fields where signal processing and Machine Learning are core techniques (e.g., Electrical Engineering and Computer Science). Though BSP may seem to be a more complicated method for estimating coding scores that can otherwise be produced with trained or naïve coding methods, BSP offers substantial potential advantages over trained and naïve coding. With regard to generating coding scores of established correlates of relationship satisfaction, the major potential advantages of BSP are that it can be used to estimate scores for as many behaviors as desired and that it can be used to estimate scores for additional behaviors at any point in time. In contrast to having to recruit and train a coding team to code additional behaviors, BSP methods allow for simply changing the task the computer is asked to perform (e.g., estimate blame scores instead of criticism scores).
In addition to these advantages, there are also several disadvantages of BSP relative to trained and naïve coding methods. First, BSP can only provide information about behavior that researchers determine a priori. While it is also true that trained and naïve coding systems are limited to the codes in their manuals, trained and naïve coders are able to provide generative feedback to researchers that could be helpful in refining a coding system. For example, coders may notice that the definition of a code that has worked well in previous research is problematic in a new application or that they are frequently noticing a behavior that seems important but that is not captured by any of the existing codes. Second, trained and naïve coders are better able to flexibly adapt to unanticipated behavioral events than is BSP. For example, if a research study asked participants to record themselves having conflict interactions at home and a couple answered a phone call during one of those recordings, a RA could easily identify that the phone call should not be considered for coding purposes while a BSP system would have a very difficult time making the same determination. Third, trained and naïve coders are better able to adapt to a wider range of recording conditions than is BSP. BSP is only as good as its ability to correctly identify the signals of interest and to extract meaning information from those signals. Background noise, particularly other speech, and low quality recordings impede BSP’s ability to perform these tasks well. With the availability and affordability of high quality audio and videorecorders, background noise and low qualities recordings are not common problems in laboratory settings. However, background noise is frequent in non-laboratory settings (e.g., public announcements in a hospital, hallway conversations in office buildings, etc.). At present, BSP would likely be ill-suited to analyzing recordings made outside of the research laboratory.

Summary. Disentangling behavioral associations for related outcomes requires inclusion of measures of relationship distress and of established behavioral correlates of relationship
distress in tests of associations between disorder/illness specific behavioral processes and disorder/illness outcomes. Established behavioral correlates of relationship distress can be coded using traditional coding methods, naïve coding methods, and BSP methods. At present, there is no clear empirical evidence demonstrating the superiority of one of these methods over the other. We recommend that decisions about which method and which coding system to use be guided by the same principles as selection of any coding system. There should be as strong a match as possible between the purpose for which the coding system was created and the purpose for which the coding system will be used in any given study. When multiple coding systems could provide a similar match with study aims, we recommend that available psychometric evidence be heavily weighted in selecting amongst them. As we will address in the next section, systems that can consistency be used to generate coding data with high interrater reliability of codes and high internal reliability of coding composites should be preferred over those that typically generate data with lower reliability values. Finally, we recommend that it is advisable to conduct tests of the psychometric properties of any naïve adaptation of a trained coding system. It is likely that many trained coding systems can be adapted, at least in part, into a naïve version but such adaptations must be carefully tested before use.

**Issue 2, Within-group comparisons**

A second set of methodological needs arise from work examining change in behavior over time. Examples of this kind of work include changes in behavior created by intervention (e.g., K. Baucom et al., 2015) and within-couple, experimental manipulation of interaction tasks (e.g., Christensen & Heavey, 1990). As is true for any association between two variables, the

---

6 An example implementation of how traditional observational coding methods, naïve coding methods, and BSP could be used in conjunction is included in the Appendix. This example additionally considers issues of interrater reliability raised in Issue 2.

© 2017, American Psychological Association. This paper is not the copy of record and may not exactly replicate the final, authoritative version of the article. Please do not copy or cite without authors permission. The final article will be available, upon publication, via its DOI: 10.1037/fam0000370
ability to detect significant change over time is hampered by data with low internal reliability. Despite widespread knowledge of the impact of low internal reliability on power to detect a significant association, the implications for detecting change over time in observational coding data are underappreciated and rarely considered. In this section, we discuss the statistical issues involved in how low interrater reliability, a form of internal reliability, impacts power to detect significant change over time and show how these issues can be incorporated into a formula (Equation 3 presented in the note for Table 1) for estimating the necessary size of a coding team to be adequately powered to detect change over time under differing circumstances.

**Statistical issues involved in detecting significant change over time.** The major statistical issues involved in the study of change of behavior over time can be understood as dividing the variance (i.e., partitioning) in behavior measured at a later point in time into three pieces: 1) variance that is shared with behavior measured at an earlier point in time (i.e., stability of behavior), 2) variance attributable to change over time, and 3) error (McArdle, 2009). Figure 2 is a series of Venn diagram style representations of this partitioning process for behavior measured twice, once at time 1 and once at time 2. The aim of these diagrams is to partition the variance in behavior measured at time 2 into variance shared with behavior at time 1, variance attributable to change from time 1 to time 2, and error. Panel A depicts the covariance between behavior measured at time 2 and behavior measured at time 1. The overlap between the two circles represents the covariance of behavior measured at the two points in time; this covariance is the stability of behavior. The portion of the variance in behavior at time 2 that is shaded in light gray is the remaining variance in behavior at time 2. This remaining variance in behavior at time 2 is the maximum amount of variance that could be attributable to change over time or error. Panel B represents the partitioning of this remaining variance in behavior at time 2 by
addition of a third circle that represents change over time. The dark gray shaded portion of the overlap between the circle for time and behavior at time 2 represents the amount of variance in behavior at time 2 that is attributable to change over time. The ratio of the dark gray shaded area in Panel B to the light gray shaded area in Panel A is the partial point-biserial correlation representing change over time (Cronbach & Furby, 1970).

Panel C depicts the impact of interrater reliability on the magnitude of the covariances (i.e., the size of the shaded areas) depicted in Panels A and B. The strength of an association between any two variables is reduced in proportion to the internal reliability of those variables so when internal reliability is less than 1.0, the observed association between variables is smaller than the true association between those variables. The implications of less than perfect interrater reliability of an observational code for detecting significant change in that behavior over time is that some of the true covariance between behavior at time 2 and behavior at time 1 is added to the residual variance in behavior at time 2 (the light gray shaded portion of the overlap between behavior at time 2 and behavior at time 1), and some of the variance attributable to change over time is added to the residual variance in behavior at time 2 (the light gray shaded portion of the overlap between behavior at time 2 and behavior at time 1). These changes in the covariance result in the proportion of the variance in behavior at time 2 attributable to change over time (the dark gray shaded area) being smaller than it should be and the residual variance (the light gray shaded areas) being larger than it should be; these two changes result in the magnitude of

7 The formula for attenuated correlation between two variables is
\[ r_{xy} = r_{x'y'} \sqrt{r_{xx}r_{yy}} \]
where \( r_{xy} \) is the observed correlation between X and Y, \( r_{x'y'} \) the true correlation between X and Y, \( r_{xx} \) the reliability of X and \( r_{yy} \) the reliability of Y (Spearman, 1904). In modeling change in a behavior, X and Y are the same thing so the observed correlation is attenuated to the extent that the reliability of the behavioral code is less than 1.

\[ r_{x@time 1 \times time 2} = r_{x@time 1 \times time 2} \sqrt{r_{xx}} \]

© 2017, American Psychological Association. This paper is not the copy of record and may not exactly replicate the final, authoritative version of the article. Please do not copy or cite without authors permission. The final article will be available, upon publication, via its DOI: 10.1037/fam0000370
observed change over time being substantially less than it truly is.

Panel C can be used to derive an equation for the partial point-biserial correlation that represents the magnitude of observed change over time (Equation 3 presented in the note for Table 1). The results of this equation can be used to estimate the sample size that would be needed to observe significant change over time depending on interrater reliability, the magnitude of the correlation between behavior measured at times 1 and 2, and the amount of error in behavior measured at time 2 (presented in Table 1). For example, assuming that 40% of the variance in behavior is attributable to change and 50% is attributable to stability (row 7 in Table 1), increasing the interrater reliability from .6 to .8 reduces the number of couples needed to detect significant change over time by approximately 40%. The pattern of change in the number of couples needed to detect significant change over time suggests that the interrater reliability of coding data should be evaluated not only for demonstrating that behavior can be reliability coded but also for the impact that the interrater reliability has on power to detect change over time.

We recommend that such evaluations be conducted not only at the manuscript review stage but also by researchers when deciding on the size of a coding team. Cronbach’s (1947) demonstration that adding more items to the measure of a construct substantially increases the internal reliability of that construct is just as true for the interrater reliability of coding data as it is for scales on a self-report measure. Table 2 presents the average interrater reliability of coding teams of size \( n \) for seven codes from the Naïve Observational Rating System (NORS; K. Baucom et al., 2012). Consistent with Cronbach’s postulation, the interrater reliability of each code (shown in the top row of each cell) increases as the size of the coding team increases. These

---

8 Coding data are for 44, 10-minute interactions collected as part of a larger study of at-risk couples during the transitioning to parenthood (K. Baucom et al., in press). All interactions were coded by 15 naïve coders.
results suggest that one simple method of increasing power to detect change over time is to create larger coding teams. Larger coding teams of trained or naïve coders would increase the interrater reliability of coding data in a manner depicted in Table 2. Naïve systems may be particularly well suited for use with large coding teams because they take significantly less time and effort to train naïve coders, and it takes coders less time to rate an interaction using a naïve system relative to a trained system (Baucom et al., 2012; Waldinger et al., 2004). Regardless of whether a trained or a naïve coding system is used, it is vital that researchers consider the impact of interrater reliability of coding data on a priori power to detect associations in future research.

Issue 3, Between-group comparisons

A third major set of issues in current observational research arises from the study of behavior across groups of couples, such as cross-cultural work or comparison of different ethnic groups on couple behavior (e.g., Hahlweg et al., 2000). Relationship science is an increasingly diverse and global science, and there is strong need to develop new/adapt existing observational coding systems for studying couples from a wide range of cultural, ethnic, and racial backgrounds. A primary challenge in developing and adapting observational coding systems for the study of different groups of couples is that groups may differ in the ways that they behave and the ways that they perceive behavior. Whether a group of couples is being studied out of interest in that particular group or multiple groups are being studied to understand similarities and differences between the groups (e.g., Peplau & Fingerhut, 2007), parsing enactment of behavior from perception of behavior is a methodological challenge that must be overcome.

Observational research on couples and families has faced a similar methodological

---

9 The Spearman-Brown prophecy formula (SBP; Brown, 1910; Spearman, 1910) provides a statistical means for estimating how many coders are needed to achieve a desired interrater reliability for coding systems where psychometric information is available. Details of using the SBP for this purpose are provided in the Appendix.
challenge previously albeit in a different form. Observational research in the 1980’s and ‘90’s explored the similarities and differences of coding data generated by trained coders and by study participants when watching recordings of themselves interacting with their partners. Modest correlations were generally found for these two kinds of coding data (e.g., Birchler et al., 1984). It is likely that participants’ subjective biases in the perception of their own and their partner’s behaviors added unique forms of variance for each couple and that these couple-specific forms of variance downwardly biased associations with data created by trained coders. At the same time, partners’ subjective biases provide unique insight into how they perceive and experience their relationships. Partner created data can be understood as representing an “insider” perspective while data created by trained coders can be understood as representing an “outsider” perspective.

A similar concept arises in work examining similarities and differences between groups from different ethnic, racial, or cultural backgrounds. The terms emic and etic refer to “insider” and “outsider” perspectives, respectively. An emic approach to measurement is “[group]-specific and … concerned with the nuanced meaning of a construct as described by representative informants.” An etic approach “[aims] to measure the universality of constructs” (Tamis-LeMonda et al., 2008). Video-recall procedures that ask romantic partners to rate their own interactions are emic while using trained coders to objectively rate couple interactions is etic.

The prevalence of studies that use trained coders relative to those that use video recall procedures suggests that etic approaches are preferred in the field. Consistent with this notion, a widely adopted operationalization of observational coding in couple and family research is to “[apply] the same, rather than idiographic, definitions and measurement of constructs to each family within the sample (Margolin et al., 1998, p.196).” There are likely many reasons for preferring trained coders such as the perception that the data they generate is more scientific.
since it is not subject to idiosyncratic cognitive biases that vary between couples, the desire to compare results across studies, and the logistic complications of video recall procedures.

For the purposes of comparing groups, emic and etic approaches to observational coding have different relative merits. Rather than one being methodologically stronger than the other, we see the two methods as generating complimentary forms of information that could potentially be used to understand behavioral pattern and associations with other variables more completely than is possible with either method alone. The thing that is most important about the distinction between emic and etic approaches to measuring behavior is that they are not interchangeable. As with selecting a coding system and determining the size of a coding team, we recommend that researchers deliberately select either an emic or an etic approach to measuring behavior and that a distinction between these two approaches be made when reviewing observational literature.

If researchers wish to pursue emic measurement of behavior, several factors should be taken into account. One factor is the background characteristics of coders. There has long been recognition that individual differences between coders likely impact the way that they rate behavior. For example, Margolin and colleagues (Margolin et al., 1998) noted, “One reality of coding is that coders bring their own characteristics and backgrounds to the task. Three obvious characteristics are the coders' gender, ethnicity, and life experience (p. 204).” Heyman (2001) likewise commented, “"Healthy" couple behavior is undoubtedly culturally determined (p. 6).”

Research on parent-child interaction has explored this possibility in depth, and the conclusions offer valuable guidelines for emic observational research on couples. Emic parent-child research uses variations on a common study design where the variable that defines the group difference in parent-child dyads is used to recruit groups of coders who differ from each other in the same way as the parent-child dyads differ (e.g., African American and European
American mother-infant dyads coded by both African American and European American coding teams; Campione-Barr & Smetana, 2004; Costigan, Bardina, Cauce, Kim, & Latendresse, 2006). The majority of these studies find evidence of an interaction between group membership of the parent-child dyad and group membership of the coding teams. There are inconsistencies in the particular form of the difference, but most findings support the existence of ethnocentric bias (e.g., Harvey et al., 2009). The most consistent evidence suggests that majority group coders (e.g., European American coders in the US) view minority group dyads more negatively/less positively than they view majority group dyads (e.g., Wang et al., 2007; Yasui & Dishion, 2008), though there is some evidence that ethnocentric bias is a more generalized phenomenon (Harvey et al., 2009). Based on these findings, we recommend that researchers who wish to conduct emic observational research with couples consider recruiting more than one coding team and that the demographic characteristics of the coding teams vary such that the members of one coding team are demographically similar to the group of couples being coded and the members of the other coding team are demographic dissimilar to the group of couples being coded but demographically similar to one another. A significant difference in the mean level of a code for the two coding teams would provide evidence of group differences in the way that that code was perceived and imply that there is significant cultural variation in the perception of that behavior.

A second factor that should be considered in future emic observational research on couples is the use of trained versus naïve coding systems. Egocentric bias has been found to be stronger in coding data created by minimally untrained coders relative to that created by trained coders. Furthermore, additional training in a coding system has been found to reduce egocentric bias (Yasui & Dishion, 2008). These finding suggests that if researchers wish to conduct emic observational measurement of couples, naïve coding systems and minimally trained coders
should be preferred over trained coding systems and highly trained coders.

Future etic observational research on couples would be well served by continuing to implement longstanding recommendations for recruiting coding teams that represent a diversity of demographic characteristics (e.g., Margolin et al., 1998) and by preferring trained coding systems over naïve coding systems. Such work would also benefit from using recommended procedures for establishing factorial invariance between the groups being compared (e.g., Little, 1997). While naïve coding systems appear to be able to generate etic coding data when the naïve coding team is large and members of the coding team are diverse with respect to sociocultural and demographic characteristics (e.g., K. Baucom et al., 2012), methods for generating etic coding data with trained coding systems and diverse coding teams are more established at present and, therefore, more advisable until additional research on generating etic coding data with naïve coding systems has been conducted.

Conclusions

The wide range of questions being examined in ongoing observational research on couples promises to pave the way for important discoveries that will improve the field’s understanding of basic behavioral phenomena and that will contribute to improving the efficacy of couple-based interventions. The contributions of future observational findings would likely be maximized by consideration of the specific and overarching issues outlined above. Rather than relying on generally agreed upon standards for observational coding, we encourage researchers to carefully consider how decisions related to which behaviors to code, what kind of coding system to use, and how to create a coding team increase or decrease the fit between the study’s

10 Establishing factorial invariance involves testing for differences in several aspects of the factor structure to increase confidence that a coding system measures the same constructs across the two groups.

© 2017, American Psychological Association. This paper is not the copy of record and may not exactly replicate the final, authoritative version of the article. Please do not copy or cite without authors permission. The final article will be available, upon publication, via its DOI: 10.1037/fam0000370
research questions and the methods used to create data for testing those questions. Fit should be evaluated in terms of ability to precisely test the primary question of interest while ruling out alternative hypotheses and ensuring adequate power for testing study hypotheses. Evaluating these elements of fit between research question and study methods is by no means new, but such evaluation has commonly fallen by the wayside in observational research on couples. We hope that the conceptual issues and statistical methods for evaluating fit presented in this manuscript will spur a renewed appreciation for these elements of design decisions in future research.
References


© 2017, American Psychological Association. This paper is not the copy of record and may not exactly replicate the final, authoritative version of the article. Please do not copy or cite without authors permission. The final article will be available, upon publication, via its DOI: 10.1037/fam0000370


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*, © 2017, American Psychological Association. This paper is not the copy of record and may not exactly replicate the final, authoritative version of the article. Please do not copy or cite without authors permission. The final article will be available, upon publication, via its DOI: 10.1037/fam0000370
doi:10.1037/0022-3514.59.1.73


© 2017, American Psychological Association. This paper is not the copy of record and may not exactly replicate the final, authoritative version of the article. Please do not copy or cite without authors permission. The final article will be available, upon publication, via its DOI: 10.1037/fam0000370


© 2017, American Psychological Association. This paper is not the copy of record and may not exactly replicate the final, authoritative version of the article. Please do not copy or cite without authors permission. The final article will be available, upon publication, via its DOI: 10.1037/fam0000370


Little, T.D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data:

© 2017, American Psychological Association. This paper is not the copy of record and may not exactly replicate the final, authoritative version of the article. Please do not copy or cite without authors permission. The final article will be available, upon publication, via its DOI: 10.1037/fam0000370


Roberts, N. A., Leonard, R. C., Butler, E. A., Levenson, R. W., & Kanter, J. W. (2013). Job © 2017, American Psychological Association. This paper is not the copy of record and may not exactly replicate the final, authoritative version of the article. Please do not copy or cite without authors permission. The final article will be available, upon publication, via its DOI: 10.1037/fam0000370


© 2017, American Psychological Association. This paper is not the copy of record and may not exactly replicate the final, authoritative version of the article. Please do not copy or cite without authors permission. The final article will be available, upon publication, via its DOI: 10.1037/fam0000370

doi:10.1080/15295190802612599


© 2017, American Psychological Association. This paper is not the copy of record and may not exactly replicate the final, authoritative version of the article. Please do not copy or cite without authors permission. The final article will be available, upon publication, via its DOI: 10.1037/fam0000370