Behavioral Observation and Coding

Richard E. Heyman

Michael F. Lorber

New York University

J. Mark Eddy

University of Washington

Tessa V. West

New York University

Chapter to appear in Harry T. Reis and Charles M. Judd (Eds.), *Handbook of Research Methods in Social and Personality Psychology (2nd ed.)*. New York: Cambridge University Press.

Author Notes

Preparation of this chapter was supported by National Institute of Dental and Craniofacial Research grant R21DE01953701A1 and National Institute of Child Health and Human Development grant R01 HD054880. We are indebted to Ashley Dills for her masterful copyediting.

Contact information

Richard E. Heyman, Ph.D.
Professor
Family Translational Research Group
Department of Cariology and Comprehensive Care
New York University
345 East 24th Street, 2S-VA
New York, NY 10010
Email: Richard.Heyman@NYU.edu
212-998-9984

Behavioral Observation and Coding

Kurt Lewin (1951, p. 169) wrote that there is “nothing so practical as a good theory.” One could add that there is nothing so practical as a good theory testing tool. We devote this chapter to one such tool — behavioral observation — that excels at both the identification of behaviors worth theorizing about and the testing of theories of behavior. This chapter provides an overview of behavioral observation, including the contexts researchers use when observing, the forms in which they record behaviors for analysis (e.g., coding), the methods available to document that different observers coded behaviors similarly (i.e., interrater agreement, an element of reliability), the necessity of establishing other forms of reliability as well as validity, and methods of analyzing behavioral observation data.

What is Behavioral Observation?

The observation of behavior is at the center of all scientific inquiry in social and personality psychology. Although there are a wide variety of methods that researchers use when observing, the term “behavioral observation” generally refers to a researcher seeing and/or hearing, and then systematically recording, the behaviors of an individual or group of individuals within a particular social context of interest, such as the classroom, the playground, the peer group, the home, the clinic, or the workplace. Typically, individuals are observed for relatively brief periods of time, but often for multiple bouts.

Sometimes observations are conducted “live.” More often, an audio or video recording is made (and sometimes transcribed into written form as well); observations are then conducted using one or more of these at the convenience of the researcher. During an observation, a researcher periodically summarizes the physical and/or verbal behaviors of the participants of interest into specific categories using a clearly defined system of “codes” that are assigned based

on a set of rules. Each code is used to mark the occurrence of a specific behavior or set of behaviors, and depending on the data collection technique, may be recorded in parallel with other information about the code (e.g., affective quality, start and stop times). The final result is a sequential record of the behaviors of one or more individuals.

Why Use This Research Method?

Some value behavioral observation data simply because it provides objective information about the frequencies of particular behaviors engaged in by a given individual. This might be important, for example, to a researcher interested in examining whether an intervention changed the frequency of certain targeted behaviors (e.g., factors that promote or inhibit bystander intervention; Penner, Dovidio, Piliavin, & Schroeder, 2005). Others value observation as a means to examine the relations between and among behaviors, either within individuals or among dyads or groups. This might be important, for example, to a researcher who is interested in whether the same behaviors are expressed during inter- and intra-racial interactions, and whether perceivers apply the same meaning to these behaviors as a function of the racial composition of the interacting dyad (e.g., Gray, Mendes, & Denny-Brown, 2008). Thus, behavioral observations are useful for answering questions not only about individual *outcomes* but also about social interactional *processes*.

In many studies, outcomes and processes are measured through self-reports from one informant. The generalizability of findings from this measurement strategy may be limited. In an attempt to rectify this problem, a high value has been placed in recent years on the use of multiple informants and multiple assessment methods to measure psychological constructs of importance (e.g., Smith & Harris, 2006; see also Brewer & Crano, ch. 2). This approach is thought to lead to a more reliable and valid index of the “true score” of a construct. Introducing

multiple informants into assessment batteries is relatively easy—various forms of self-report questionnaires on behaviors or behavioral patterns are readily available, or can be created relatively easily, for different reporters (e.g., parents, teachers, youth).

Although behavioral observation is a quite appealing method for some researchers, it does have its downsides. Even if an existing coding system is identified for a new study, purchasing the necessary equipment, securing private coding space, and assembling and training a team of observers (i.e., a "coding team") can be time consuming and expensive. Once a team is ready, collecting data *in vivo*, or collecting and storing video or audio records and transcribing those records, and then managing and analyzing the resulting data, can also be quite costly.

Furthermore, although the focus of a typical coding team is usually on obtaining and maintaining interrater agreement (i.e., independent observers applying the same codes to a given stream of behavior), this is no guarantee that behavioral observation will generate reliable (i.e., stable) or valid (i.e., "true" measures) scores of constructs of interest in a given sample. Indicators derived from behavioral observation often are weakly correlated with self-report measures of the same constructs, and the meaning of this may be unclear. Finally, the existence of audio or video records creates ongoing human subjects issues related to the protection of confidentiality and anonymity. In short, despite their appeal, "observational data, compared with other forms of data, are unwieldy and messy" (Margolin et al., 1998; p. 29). Nevertheless, behavioral observation has been employed frequently over the past 50 years, particularly among psychologists interested in interpersonal and intergroup relations, human development, and close relationships.

Observational Settings

Observational settings exist along a continuum of researcher influence ranging from

unfettered natural environments to tightly controlled experimental situations. Purely naturalistic situations have the advantage of being high in ecological validity (see Cialdini & Levy, ch. x). Although researchers observing behavior in its natural environment still need to establish the reliability of their observations (e.g., consistency across observers, episodes, or settings), the real world generalizability of such observations is self-evident. The more the researcher intervenes in the setting to be observed, the more has to be done to demonstrate that the setting produces externally valid results.

In the sections below, we provide an overview of different degrees of researcher interventions into settings. As with any research tool, the validity of behavioral observation is situation-dependent and can only be inferred from that tested, narrow use; it is not "proven" for all time (e.g., Haynes & O'Brien, 2000). Thus, behavioral observation cannot be said to be a valid assessment approach any more than questionnaires can be said to be a valid assessment approach.

Naturalistic Observation

Naturalistic observation has a long history in the study of animal (e.g., Lorenz, 1970, 1971) and human (e.g., Mead, 1928) behavior. Some researchers who favor this type of observation use a qualitative approach, where the coding system is not predetermined. Others use a quantitative approach, marked by the use of preset codes and precisely defined rules for their assignment.

One of the most important studies in social psychology—~~Festinger~~, Riecken, and Schacter's (1956) *When Prophecy Fails*, which focused on social interactions within a doomsday cult and proposed cognitive dissonance theory—used naturalistic observation. Observers ultimately were not outsiders, but rather became members of the social group being observed. There were no predetermined codes to classify behaviors. The observers who infiltrated the cult

received only an overview of the study's purpose and the phenomena of highest interest. As Festinger et al. (1956, p. 248) described, "Problems of rigor and systematization in observation took a back seat in the hurly-burly of simply trying to keep up with a movement that often seemed to us to be ruled by whimsy." The researchers also noted a common problem with many naturalistic studies: "[Observing] was frequently irritating because of the irrelevancies… that occupied vast quantities of time …. [and] the repetitiousness of much that was said." Observing surreptitiously without modern hidden recorders also necessitated observers taking frequent bathroom breaks or walks outdoors to absent themselves from the group to take notes.

Whereas Festinger et al. used naturalistic observation to examine an extraordinary social situation, most investigators use this approach to examine the ordinary (i.e., how interactions during normal life are related to particular outcomes of interest). To facilitate the collection of this type of information, Mehl, Pennebaker and colleagues (e.g., Mehl, Pennebaker, Crow, Dabbs, & Price, 2001) developed a behavior observational paradigm that employs the Electronically Activated Recorder (EAR; Mehl, 2007; Mehl & Robbins, 2012). The EAR is an audio recorder that is worn in everyday settings and is programmed to make 30 second audio samples every 12.5 minutes (i.e., five minutes of recordings per hour).

The EAR has been used to examine questions such as "Do women really talk more than men?" (From the research done so far, it appears that they do not; e.g., Mehl, et al., 2007). The coding system developed for this paradigm, the Social Environment Coding of Sound Inventory (SECSI; Mehl & Pennebaker, 2003; Mehl et al., 2006), comprises four categories (with codes within those categories): (a) current location (e.g., home, outside), (b) activity (e.g., watching TV, eating), (c) social interaction (e.g., alone, talking on phone, in group), and (d) behavioral indicators of mood (e.g., laughing, crying, arguing). Although this work has produced important

findings related to health, Mehl (2007, p. 370), drew the same conclusion as Festinger on the banality of observing life naturalistically: “One of my first ‘aha!’ experiences when we started doing EAR research was how ordinary and mundane real life really is. The sound files we obtained from participants first and foremost documented that for most people most of real life is not thrilling, glittery, and extraordinary.”

Another recent use of naturalistic observation was of families of dual-earning parents in California (Campos, Graesch, Repetti, Bradbury, & Ochs, 2009). Because the investigators had an overwhelming 35 hours of video from two weekdays per family, Campos et al. (2009) focused only on the two minutes captured when the partners reunited after their workdays and coded these simply (i.e., positive, negative, ignoring/distracted, reporting information, checking in about logistics). The authors also presented data from the “scan sampling” of family interactions, in which, every 10 minutes, observers noted the location of each family member. They found that working couples spend almost no time together without children. In later analyses, they found that men’s, but women’s, “neuroticism” (i.e., temperamental negativity) moderated the relationship between job stress and at-home behavior (Wang, Repetti, & Campos, 2011). For instance, men high in job stress but low in neuroticism were more socially withdrawn during their first hour home, but their interactions with their children were more intense.

Quasi-Naturalistic Observation

As implied by the Festinger and Mehl quotes, naturalistic observation often requires so much time that it is inefficient and impractical. Thus, observation typically occurs in situations that are not completely natural and uninfluenced by the investigator. When investigators use quasi-naturalistic observations, the generalizability of behavior is of the highest concern and investigators attempt to influence the situation as little as possible.

The work of the Oregon Social Learning Center (OSLC) research team (e.g., Reid et al., 2002) is a model of the development and refinement of a quasi-naturalistic observational paradigm. Starting in the late 1960s, OSLC researchers wanted to conduct naturalistic behavior observations of families but quickly learned that the natural world was not conducive to cost-effective data collection (Patterson, 1982). Family members typically disappeared or sat transfixed in front of a video screen when observers arrived (and this was usually a solitary television screen, long before the advent of other screen-related distractions in the home, such as smart phones, iPads, computers, video games, etc.). Out of necessity, eight rules (see Table 1) were imposed on families during their in-home observation sessions. Patterson (1982) noted that the rules transformed the otherwise typical environment into something close to, but not identical to, the real world (i.e., those being observed were unnaturally constrained but otherwise acting naturally in their natural environment). This increases the quality of the data collected by increasing interaction but reduces generalizability slightly, exactly the kind of trade off that all researchers must weigh in designing protocols.

OSLC developed its quasi-naturalistic paradigm through trial and error, guided by both the empirical literature and by their theoretical model. The researchers were most interested in children's aversive and aggressive behaviors and their parents' responses to these behaviors. To increase the chance of observing such interactions, dinnertime was chosen as the setting to observe because earlier studies had found that mothers reported the most conflict with their children during the time surrounding meals (e.g., Goodenough, 1931). The further limitation of distractions increased the likelihood that the observational sessions would generate enough conflict behavior for hypothesis testing. Next, they tested observer influences on the data to identify whether any adjustments to their protocol were needed (e.g., they examined if

"warm-up" sessions were necessary for families to adjust to having observers in their homes; results indicated that such accommodations were unnecessary, Patterson, 1982; see also Thornberry & Brestan-Knight, 2011). They examined the frequency of key behaviors and determined how much observation over how many sessions were needed to get a stable index of the behaviors of interest. They found that 60-100 minutes of data sampled in five-minute blocks over the course of several sessions provided minimally stable estimates of boys' coercive behaviors. Finally, by using observations in a multi-trait, multi-method assessment strategy (e.g., parent reports, global observer impressions, and school or arrest data), OSLC provided evidence for the validity of their observational approach and their coding system (e.g., Patterson, Reid, & Dishion, 1992).

Analogue Observation

Although naturalistic observation might be appealing because the required inferences about generalizability are minimized, analogue situations are often preferable because of their efficiency. Social psychologists employ analogue situations to (a) create environments where otherwise difficult or impossible to observe behaviors occur (e.g., how positions of power can evoke degradation, Haney, Banks, & Zimbardo, 1973); (b) enable observation of dynamic qualities of social interaction (e.g., escalation and de-escalation of negativity in mother-child dyads; Snyder, Edwards, McGraw, Kilgore, & Holton, 1994); and/or (c) isolate determinants of behavior.

An example of an observational paradigm of this type is the couple problem-solving discussion (Heyman, 2001). Investigators typically ask couples to discuss one or two potential conflict areas for 10 to 15 minutes each. Within these general parameters, there is wide variability in exactly how conversations are structured. A prototypical protocol is presented in

Table 2. Other approaches include providing couples with standardized topics to role play (e.g., planning a vacation) that may not relate to their own conflicts (e.g., Aron, Norman, Aron, McKenna, & Heyman, 2000) or having them reenact prior conflicts (e.g., Margolin, Burman, & John, 1989). Other researchers have set up situations to observe couples providing social support (e.g., Pasch & Bradbury, 1998), sharing exciting activities (e.g., Aron et al., 2000), or discussing situations of high import (e.g., Schmaling, Wamboldt, Telford, Newman, Hops, & Eddy, 1996).

Perhaps surprisingly, asking couples to engage in communication about conflictual topics while researchers watch tends to elicit behavior with reasonable external validity. First, observed conflict behaviors in home and laboratory settings tend to be similar, although lab conflicts are a bit less negative (e.g., Gottman, 1979; Gottman & Krokoff, 1989). Second, couples judge in-lab behavior as typical of what they do at home (Foster et al., 1997). Third, partners' reactivity and self-consciousness while being observed are relatively low (Christensen & Hazzard, 1983; Jacob, Tennenbaum, Seilhamer, Bargiel, & Sharon, 1994). Thus, even if in-lab "conflicts on command" are not quite as negative as they are at home, they still reveal detectable differences in affect, behavior, physiology, and interactional patterns and processes (e.g., Gottman, 1979, 1994, 1999).

Experimental Manipulation

Social psychologists study behavior within controlled laboratory settings to (a) observe behaviors that are not likely to be observed in unstructured settings and/or (b) to experimentally manipulate the causes of those behaviors. By controlling all aspects of a laboratory environment except that which is being manipulated, psychologists are able isolate particular behaviors of interest and make conclusions about the case of behaviors—an integral step to theory development (see Smith, ch. 3). In addition, often in naturalistic settings there are multiple causes of behaviors that are interdependent, making it difficult to isolate which of several factors

actually cause the behavior. With experimental manipulation, researchers can tease apart these causes by systematically manipulating them.

There are several issues to consider when designing an experiment in which the goal is to change behavior. Whether the manipulation is minimal or large and the degree to which behaviors are "difficult or easy to influence" are important considerations (Prentice & Miller, 1992; p. 162), and they are certainly relevant for studies that intend to influence the display of dynamic, interpersonal behaviors. Minimal manipulations that have large effects on behaviors can be particularly convincing in demonstrating the strength and size of an effect. The mere exposure effect and the minimal group paradigm are classic examples of minimal manipulations that produce large effects on behavior. As a more recent example, Goff, Steele, and Davies (2008) demonstrated that White participants who were led to believe that they would discuss racial profiling with an African American participant placed their chairs farther apart from their partners' chairs than did Whites who were led to believe that they would discuss a race-neutral topic. Goff et al.'s (2008) manipulation is minimal because the mere belief that participants would have a race-based discussion was sufficient to alter behavior.

It is also important to consider the behaviors that are manipulated and measured. It is provocative to demonstrate that an experimental manipulation affects behaviors that are "difficult to influence" (Prentice & Miller, 1992, p. 162), largely because easy to influence behaviors are mundane (e.g., ask participants to sit when they arrive and they sit is of little interest). Rapport building within cross-race interactions and conformity to groups (e.g., Asch, 1951) are examples of difficult to influence behaviors. Manipulations that are both minimal in nature *and* exert effects on such behaviors are often deemed particularly impressive by social psychologists, and are therefore more likely to make a scientific impact.

Studies that examine dynamic interpersonal behaviors, such as mimicry (Van Baaren, Janssen, Chartran, & Dijksterhuis, 2009), self-disclosure (Altman & Taylor, 1973), or rapport-building (Tickle-Degnen & Rosenthal, 1990), require at least two individuals. As such, one of the most important methodological choices that social psychologists make in terms of experimental manipulations within social interaction studies is whether or not to use a confederate, rather than another participant, as the social interaction partner of interest. If the theoretical question of interest is interpersonal in nature—that is, it involves manipulating and measuring the behaviors of both partners within the interaction or examining the interdependence between partners' behaviors—one should strive to design a study in which real participants are used, not confederates. However, there are a variety of situations where confederates may be appealing.

First, confederates offer a great deal of experimental control within an interpersonal interaction and are ideal in examining theoretical questions that are intrapersonal in nature. For example, Lakin, Chartrand, and Arkin (2008) examined how being socially excluded prior to a dyadic interaction influenced mimicry of the interaction partners' nonverbal behaviors. After receiving a social exclusion manipulation, participants interacted with a confederate who was trained to engage in a specific set of nonverbal behaviors, namely foot wiggling. The authors were interested in the degree to which participants who were socially excluded also wiggled their feet. In this example, the empirical question was *intrapersonal*—it involved examining how an individual-level predictor (social exclusion) influenced the behaviors of only one person in the interaction, not both. Second, confederates are a valid choice when the interaction partners' behaviors are the experimental manipulation. For example, Blascovich, Mendes, Hunter, Lickel, Kowai-Bell (2001) had participants interact with a stigmatized other who had a large birthmark

on her face—which was painted on using make-up—or no birthmark. The presence of the birthmark was the experimental manipulation. Third, confederates allow for a clean standardization of the dependent behavior of interest. In Lakin et al. (2008), for example, mimicry was clearly (and simply) defined as foot-wiggling. Fourth, because confederates offer a level of experimental control that participant interaction partners do not, they allow researchers to isolate the causes of behavior to gain a better understanding of social processes.

There are a few important steps that must be taken when using confederates as social interaction partners. First, it is important to make sure that confederates are not a hidden source of variance. For pragmatic purposes, researchers often use two or more confederates in a study. These confederates might not always behave consistently with each other, so one must make sure that the effect of the experimental manipulation does not depend on with which confederate participants interact. One potential method for addressing this issue is to treat confederate (e.g., Amy the confederate versus Stacy the confederate) as a predictor of the dependent behavior of interest and as a moderator of the effect of condition on that behavior (making sure that the confederate is crossed with condition). Confederate can also be considered a source of variance in the analysis. Kenny, Mohr, and Levesque (2001) discuss methods for examining reliability of observers' judgments of participants' behaviors, many of which are applicable to studies that use confederates. For example, they discuss the importance of treating the observer as a source of variance— a method that can be easily adapted to treating the confederate as source of variance.

Second, whenever possible, confederates should be blind to condition so that their behaviors are not inadvertently influenced. For example, Mendes et al. (2008) went to great lengths to ensure that the confederate did not know whether she had a birthmark painted on her face. Third, confederates may be trained to behave in a certain, consistent way across

participants, but they might engage in automatic behaviors that are outside of their awareness, especially during social interactions, and these behaviors could influence the interaction. To make sure that confederates behave consistently across participants and across conditions, researchers should record the behaviors of confederates within each interaction if possible; for example, by videotaping them and then coding their behaviors.

Sometimes confederates are used because they represent groups that are difficult to recruit to participate in research, either because they are not part of a convenience sample, or because they are a small percentage of the sample population. In these cases, confederates serve a pragmatic purpose, even when the question of interest is interpersonal. For example, many cross-race interaction studies conducted in the United States have recruited White participants who then interact with African American confederates. Although such a strategy allows the examination of cross-race encounters within the lab, this strategy limits the understanding of cross-race interactions from the African American perspective (Shelton & Richeson, 2006). As such, theories about the nature of cross-race interactions have become "one-sided" in that there is much cumulative knowledge about the attitudes and behaviors of Whites but much less knowledge about the attitudes and behaviors of African Americans. This is just one example of how the use of confederates can have direct, and potentially profound, theoretical implications.

Behavioral Observation Coding Systems

Behavioral observation coding systems tend to be one of two types. *Topographical* coding systems measure the occurrence of behaviors. *Dimensional* coding systems measure the intensity of behaviors along a dimension (e.g., warmth, engagement). The choice of a coding system depends on the specific purposes of a study. Because of the costs involved in launching a behavioral observation enterprise, a system should be only as complicated as is necessary to

fulfill the purposes of the research study. Ideally, one can find an existing system that meets the researcher's needs. As Bakeman and Gottman (1997, p. 15) noted, however, this choice should not be taken lightly: "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. [Using] 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."

Consequently, the researcher should begin with a set of hypotheses and design the coding system around these hypotheses. It is unfortunate when researchers realize *after* the coding has been completed that they failed to code a critical behavior. Given the need for specificity and completeness, a system should not be chosen without a researcher spending a significant amount of time observing and studying the phenomena of interest in a variety of ways. For example, a project might begin with a researcher watching and making observations with written or verbal notes over a period of several months. During the same period of time, a literature review can be conducted to find similar projects and to learn about what coding systems were used and how various practical issues were handled. Considerations of interest during this period of the research project include not only what to code, but also when, in what settings, and by whom. As ideas narrow, pilot work will be required with practice participants and design modifications and changes will likely follow.

This background work might lead a researcher to discover that a "just right" coding system is simply not available. In this case, the researcher is in good company. Many coding systems are derivatives of past systems that were deemed in need of revision for various reasons. For example, the Family Interaction Coding System (Patterson, 1982) was developed during the 1960's for coding naturalistic family interactions in the home setting. This system was soon

revised into the Marital Interaction Coding System (Weiss & Summers, 1983), which was developed for coding couples problem solving interactions in a laboratory setting. Like Latin, these two “dead coding languages” are the source for dozens of offshoots (Kerig & Baucom, 2004; Kerig & Lindahl, 2000). Thus, a first step in developing a new system is to try to find a past system that is closest to what is needed and revise from there. The advantage of using an existing coding system (or a close derivative of one) is that much psychometric work on reliability, interrater agreement, and validity has already been conducted. The disadvantage, as just noted, is that existing coding systems might not be a good match for one’s hypotheses.

Coding Units

The most fundamental property of a coding system is the sampling strategy for behavior, otherwise known as the “coding unit.” Coding units divide an observation into discrete segments, and each segment has the opportunity to be assigned a code, should one apply. The major sampling strategies employed in behavioral observation (see Table 3) are event, duration, interval, and time. Each strategy yields a different type of coding unit. Advantages and disadvantages of each of these strategies are discussed in Bakeman and Gottman (1997) and Haynes and O’Brien (2000). With each strategy, data richness and quality (e.g., retaining the sequential unfolding of events, reliability, validity) must be weighed against practical issues (e.g., expense, time, availability or practicality of recording devices, difficulty obtaining reliability). As noted by Margolin et al. (1998), even when a coding unit is presumably clear, technical issues, such as the quality of the audio track on a video recording or the speed of turn taking in an interaction, can make detecting some units difficult. This is one reason why researchers interested in verbal communication often create written transcriptions that are used together with audio and video feeds when coding.

Molar Versus Molecular Approach

Another key property of a coding system is how often codes are recorded. In molar, or "global," coding systems (e.g., Rapid Couples Interaction Scoring System; Krokoff, Gottman, & Haas, 1989) summary ratings are made for each code over a large number of potential coding units (e.g., every three minutes in a 15 minute observation, or once at the end of the observation). Codes tend to be few, representing behavioral classes (e.g., negativity, attentiveness, escalation, reciprocation). Thus, numerous examples of the codes of interest may occur within multiple potential coding units, but only one summary score is given, usually indicating the frequency with which a code appeared throughout the observation period.

In contrast, molecular, or "microbehavioral," systems code behavior as it unfolds over time and tend to have many fine-grained codes (e.g., eye contact, criticize, whine, withdraw) that are given within each coding unit. The large number of codes in many microbehavioral systems may make them inefficient to use, even with highly trained coders. This is because (a) coders can almost never get or maintain adequate inter-rater agreement on such a large number or codes; and (b) the codes occur too infrequently in a limited observational period to make them all useful even if they were reliably coded. Thus, researchers often resort to grouping codes, often condensing down a large system into positive, negative, and neutral classes for analytic purposes (see review in Heyman, 2001). Imagine spending the extreme time and expense required to train coders on 40 codes, only to end up only analyzing positive, negative, and neutral!

Microbehavioral systems tend to be topographical; global systems can be either topographical or dimensional, though dimensional coding, especially on a behavior-by-behavior basis, is less common. Given that many theoretical models of interest have implicit or explicit intensity X time predictions (e.g., Patterson's [1982] Coercive Family Process model posits that

reinforcement of escalating negativity contributes to the development of antisocial behavior in boys), this is unfortunate.

Noting drawbacks of microbehavioral coding systems (e.g., time to code and train, need to combine micro codes into categories, difficulty achieving interrater agreement) but wanting to retain the advantages (e.g., specificity, sequential relations), researchers (e.g., Gottman, 1996; Heyman, 2004) began developing a new generation of coding systems that contained codes which could be analyzed without resorting to massive agglomeration (e.g., categories such as "hostility" instead of separate codes for negative voice tone, hostile content, eye rolls, etc.). Some of this work was guided through statistical analyses rather than a priori decisions. For example, Heyman, Weiss, Eddy, and Vivian (1995) factor-analyzed observations of over 1,000 couples that had been coded with a 40 code system to derive a system that could code at a categorical level, thus streamlining training, coding, and analysis. The resulting system (Heyman, 2004) was still microanalytic but was much more practical to use (and had better reliability and validity) than the coding system it replaced (see also Whaley, Pinto, & Sigman, 1999).

Global systems are simpler and faster and can sometimes represent the construct of interest better (e.g., an overarching construct such as overreactive parenting may be better coded with a global code, where context can better be taken into account, than with microbehavioral coding). Some constructs, such as progress made in a problem solving task, can only be coded globally. However, agreement can sometimes be difficult to obtain due to the lack of anchoring of ratings to specific behaviors. Furthermore, global systems do not maintain sequential relations, making them less useful for analyzing patterns (unless the coders specifically coded for that pattern). In an effort to obtain the "best of all worlds," microanalytic and global systems

have been paired, usually by asking coders to make global impressions ratings after coding microanalytically (e.g., Patterson, Reid, & Dishion, 1992).

Multiple Dimensions

Another property of a coding system is how many different dimensions of an interaction are coded, and how many different codes are included within each dimension. For example, some coding systems record information about the general context within which a behavior is occurring (e.g., in a system focused on child behavior at school, the location of an interaction, such as on the playground, in the lunchroom, or in the classroom), as well as the specific behaviors of interest. Other systems might also include a code describing the quality of the behavior, such as whether it was delivered with negative, positive, or neutral affect. The choice of how many dimensions to code depends on the specific hypothesis of interest, but issues can get confused in no small part because of the high cost of conducting observational work. Once data have been collected and a team has been assembled, it may seem appealing to collect as much information as possible while coding so that a variety of tasks can be accomplished, from hypothesis testing to hypothesis development. The most obvious risk in such an approach is increased difficulty in reaching an acceptable level of interrater agreement, but it may overly burden coders and compromise even more important qualities, such as the reliability and/or validity of the observation. This can only be known if other types of data (from multiple sessions, from multiple informants, through multiple methods) are collected to aid in understanding the observational data that is collected.

Example

An example of a mature coding system is the Interpersonal Process Code (IPC; Rusby, Estes, & Dishion, 1991), a distant tributary of the aforementioned Family Interaction Coding

System. In the IPC, a target individual is chosen as the focus of an observation, and everything that individual does, and has done to him/her, is coded. The coding unit is a codeable behavior, which can continue even when the behaviors of others are also taking place (e.g., a target child starts to hum and continues to hum, even though the child he is playing with is yelling at him). When no codeable behavior is occurring, a Stop code is entered. When an individual cannot be fully heard or seen, an Out of View code is given. The IPC is coded on a handheld or stationary computer in real time and has been used to code both live and videotaped sessions.

Three dimensions are coded simultaneously in the IPC: Activity, Content, and Valence. Activity refers to the general context within which a social interaction is taking place and varies depending on the study. An example of Activity codes used in prior studies are Work, Play, Read, Eat, Attend, or Unspecified. Activity codes are given in priority so that if a code with theoretically higher priority occurs, it is given (e.g., Work trumps Play, Play trumps Read, etc.). Content refers to specific behaviors of interest. Thirteen Content codes constitute the IPC, and include positive, neutral and negative verbal, non-verbal, and physical codes. For example, the code Positive Interpersonal is assigned to “verbal expressions of approval of another’s behavior, appearance or state” (p. 17). Valence refers to the emotion tone accompanying the delivery of content (i.e., Happy, Caring, Neutral, Distress, Aversive, and Sad). In addition, who displayed the behavior (the Initiator), and whom the behavior was directed toward (the Recipient), are also coded.

Training Observers

The careful training of observers (i.e., “coders”) is essential to behavioral observation. People who may have very different perceptions of behavior must, through the training process, come to be interchangeable with one another. Moreover, they must maintain consistency over

time. By analogy, two different watches should show the same time. Over time, the watch's estimates should remain unchanged. Of course, the social judgments made by human observers are known to be fallible and certainly less precise than a watch. Thus, interrater agreement should be meticulously attended to. Failure to do so can result in increased error variance, which constrains reliability and hampers the capacity to find associations of the coded behavior with other factors (even if they truly exist).

Coder training will be covered in a somewhat cursory fashion here, having been described in more detail elsewhere (e.g., Bakeman & Gottman, 1997). We will use the example of making ratings from video recordings, although the principles are broadly applicable to coding live or from audio only. The first phase of training involves familiarizing the coders with the constructs being measured and the observational context (being careful not to reveal study hypotheses) and the reasons for the heavy focus on obtaining interrater agreement. A manual should (a) describe the coding procedures in detail; (b) clearly spell out distinctions among behaviors and (c) provide illustrative examples. Because a video example is worth a thousand words, the trainer should have an ample supply of video clips illustrating the behaviors. We recommend beginning with cardinal examples that are relatively easy to discern. Over the course of training, the examples should get progressively more challenging, illustrating finer distinctions. Meetings are typically held two to three times per week with the trainer and all coders present. Between meetings, coders practice on a carefully selected set of video recordings. The training videos should be selected to illustrate the full range of behaviors being coded, with progressively more difficult cases presented over time. Meetings are used to review the process of coding (e.g., the reasoning behind coding decisions) and clarify decision rules and sources of disagreement. Interrater agreement is calculated for each video assigned and reviewed in the

meetings. This phase lasts for as long as necessary to achieve sufficient agreement. For categorical data, we suggest training coders until they consistently agree with the ratings of a master coder about 70% to 90% of the time, depending on the complexity of the coding system. For dimensional ratings, we suggest training coders until the majority of scores are in point-by-point agreement with a master coder, with disagreements very rarely greater than one point. These strategies will usually well exceed standard benchmarks for acceptable interrater agreement (i.e., *Kappa* or *AC1* above 0.6, *ICC* above 0.7; see "Interrater Agreement" section below).

On reaching the above criteria, the coders are ready to begin producing "real" data, and there is a shift from training to maintenance. In the most typical case, the videos to be coded are divvied up among coders, with only partial overlap in which videos are coded by two or more different people. These overlapping cases are used to assess interrater agreement (to be reported in resulting manuscripts); thus, it is crucial that the coders are not informed about which videos will be used for assessing agreement. Additionally, it is important that these "reliability videos" are selected at random, typically during each week of coding. After all coders have completed the reliability videos each week, they are then reviewed in meetings and the reliability statistics are presented to the coders. The purpose of these meetings is to maintain coders' performance and prevent shifts in the use of rating criteria over time. Typically, the reliability sample consists of a randomly selected 25% to 33% of all videos coded.

Interrater Agreement

Clearly, interrater agreement is an important consideration when coding. One must be able to establish that the codes recorded from an observation are not just one person's idiosyncratic view of the world, but reflect a standard, albeit imperfect, set of definitions that can

be applied with nearly identical results by other people and in the same manner across time. Interrater agreement statistics are quantitative aids for this task. They clearly are useful and vital in training coders, where interrater agreement statistics can be used to monitor progress toward a quantitative agreement criterion. Interrater agreement statistics can also be usefully employed to monitor and correct drift in the use of rating criteria once coders move beyond training. Finally, reporting of interrater agreement in published works is important to help readers evaluate a study's methods and findings.

Although "interrater agreement" and "reliability" are sometimes used interchangeably, this is sloppy usage. Mitchell (1979) provides a cogent discussion of the critical distinctions, grounded in classical measurement theory. To summarize, reliability reflects the degree to which variability in obtained scores (e.g., the ratings assigned by a coder) reflects variability in the underlying trait being measured. Interrater disagreements reflect only one of several threats to reliability; others include random fluctuations in subjects' behavior, in the setting, and in the protocol. Thus the degree of interrater agreement is not equivalent to the degree of reliability in the measure of behavior, as it is only one piece of the pie. However, because interrater agreement is controllable by the investigator, it has received the most attention. As disagreements increase, measurement error increases and reliability and validity decrease.

Which Interrater Agreement Statistic to Use

The scale of measurement and the variables formed from observational data largely determine the interrater agreement statistic used. One set of statistics is most appropriate for categorical or nominal judgments (e.g., deciding which of several emotions a person is expressing), which tend to be the basis for molecular, or microbehavioral, coding systems. Another set is more appropriate for ordinal, interval, or ratio scale judgments (e.g., deciding how

intense a person's expression of a given emotion is), which are sometimes the basis for global coding systems. The distinction is not absolute, however, depending on the intended usage of the observations. For example, categorical ratings are often summarized across an entire observation period into frequency and duration variables, in which the latter set of tools can be applied; however, Bakeman and Quera (2010) maintain that it is still important to establish behavior-by-behavior (i.e., local) agreement in these contexts. In contrast, if behavioral sequences are to be analyzed, then establishing behavior-by-behavior agreement would be required, not optional. In the following sections, we describe the most common interrater agreement statistics, as well as some useful alternatives.

Categorical Observations

Interrater agreement statistics for categorical observations each begin with the raw proportion of agreement between raters. Yet, even people who make ratings purely at random agree with one another some of the time by pure chance. Accordingly, interrater agreement statistics adjust for this possibility, with the differences among these adjustments responsible for the differences between the statistics.

The frequencies of agreement and disagreement are helpfully represented in what is known as a "confusion matrix" (see Table 4). A simple confusion matrix is presented for the situation in which two coders' agreements and disagreements in the presence vs. absence of a given behavior are represented. Agreements are found in bold along the diagonal, with "a" representing the number of agreements on the presence of a behavior, and "d" representing the number of agreements on the absence of a behavior. Disagreements are found in the off-diagonal cells, "c" and "b." The row ("e" and "f") and column ("g" and "h") totals are referred to as "marginal frequencies" or simply "marginals"; they represent the frequencies for each coder's

ratings of behavior presence and absence. Each of the interrater agreement statistics are calculated from the tallies in the confusion matrix.

Cohen's Kappa. Cohen's (1960) kappa is probably the most widely used interrater agreement statistic and is given by the following formula, with reference to the Table 4 example of a single code's presence or absence rated by two observers:

$$\kappa = (P_o - P_{e|\kappa}) / (1 - P_{e|\kappa}),$$

where P_o is the observed agreement, found along the diagonal of Table 4 and given by

$$P_0 = (a + d) / i,$$

and $P_{e|\kappa}$is the kappa model expected or chance agreement, calculated by considering the row and column marginals and given by

$$P_{e|\kappa} = [(e * g) / i + (f * h) / i] / i.$$

Kappa generalizes to accommodate multiple codes; however, code-by-code interrater agreement is essential to establish and report, as disagreements on a code can go unnoticed if embedded in a larger matrix with other codes for which there is better agreement. Moreover, kappa tends to be larger with a greater number of codes (Bakeman, Quera, McArthur, & Robinson, 1997), potentially yielding overly optimistic estimates of interrater agreement.

Kappa is straightforward to calculate (by hand or by using spreadsheets such as Excel), but it can also be calculated in standard statistics programs (e.g., SPSS) and with Robinson and Bakeman's (1998) ComKappa program; the 2010 update is available for download from Bakeman's Internet site: www2.gsu.edu/~psyrab/BakemanPrograms.htm. Kappa can also be calculated with the "irr" package in the free statistics program, R (cran.r-project.org/web/packages/irr/irr.pdf; R Development Core Team, 2005).

By far the greatest limitation of kappa is how it is affected by distributional asymmetries

(i.e., high or low rates of a given behavior). These distributional asymmetries are referred to as "skewed marginals" because the row or column totals or "marginals" ("e" vs. "f" or "g" vs. "h" in Table 4) are lopsided, as they tend to be in psychological research. This is in large part because —as noted earlier by Festinger and Mehl—many of the most interesting psychological phenomena are relatively infrequent compared with the mundane. The effect of skewed marginals can be seen in Figure 1. Panel A models the performance of Cohen's kappa, as well as the other statistics in this section given 90% interrater agreement, with evenly apportioned disagreements in the presence vs. absence of the behavior being rated. When a behavior occurs at a rate of 50%, and the marginals are thus perfectly balanced, kappa is .80 (i.e., a 10% downward adjustment for random agreement). The greater the deviation from balanced marginals, the greater the adjustment. When the behavior reaches an 80/20 split (i.e., present or absent 80% of the time), kappa is .69 (i.e., a 21% chance agreement adjustment). The adjustment is even greater when behaviors are very rare or very frequent, with kappa falling to .44 at a 90/10 split. Panel B shows that kappa's sensitivity to skewed marginals is even greater at 80% interrater agreement.

Typical rules of thumb for interpreting kappa and similar statistics are that kappas from .40 to .59 are fair, .60 to .74 are good, and .75 and above are excellent (Cicchetti, 1994). Yet the skewed marginal problem severely challenges these guidelines (Bakeman et al., 1997), as it is nearly impossible to achieve substantial kappas with highly skewed data. Gwet's (2008) Monte Carlo analyses demonstrate that kappa's chance agreement is incorrect for very common or uncommon behaviors, thus decreasing the utility of kappa in a very common research scenario.

Weighted Kappa. Weighted kappa (Cohen, 1968) is an alternative to kappa that allows the researcher to penalize more heavily for some disagreements than others. In contrast, unweighted kappa regards all disagreements as equally serious. The weighted kappa is rarely

used with nominal data in social psychology, perhaps due to the difficulties in establishing and convincing others of the validity of the weights (Bakeman & Quera, 2011). However, it involves creating a weights matrix that specifies the severity of disagreements. For example, one might decide that disagreements in rating different forms of negative emotion are less serious than disagreements in rating negative vs. positive emotions. This might lead one to weight anger vs. contempt disagreements as 1 and happiness vs. anger or contempt difference as 2 in the "weights matrix" (i.e., a grid containing the weights for all possible combinations of ratings of the two coders being compared). Agreements are assigned 0 in the weights matrix. The weights are simultaneously taken into account, alongside the observed and expected or chance agreements in calculating the weighted kappa. Weighted kappa is given as:

$$\kappa_w = 1 - \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij} x_{ij}}{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij} e_{ij}} \kappa_w = 1 - \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij} x_{ij}}{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij} e_{ij}},$$

where k is the number of codes, and w_{ij}, x_{ij}, and e_{ij} correspond to elements (i-th row and j-th column) in the weight, observed, and expected matrices, respectively. Borrowing Bakeman and Quera's (2011) notation, $e_{ij} = p_{+j} x_{i+}$ with x_{i+} the sum of the i-th row, p_{+j} the probability for the j-th column, and $p_{+j} = x_{+j} / N$.

Fortunately, weighted kappa can easily be computed using spreadsheets such as Excel, with ComKappa or with the "irr" package in R (see above).

Van Eerdewegh's V. Spitznagel and Helzer (1985) offer a statistic called *V* as an alternative to kappa that is less sensitive to the skewed marginal problem. We refer to this statistic as Van Eerdewegh's *V* (after the statistic's author), to distinguish it from Cramér's *V* (Cramér, 1946). *V* is given by:

V = [(√(b1 * c2) - √(c1 * b2)] / [√(b1 + b2) * √(c1 + c2)].

As seen in Figure 1, *V* is identical to kappa with balanced marginals, with greater differences at greater splits, in which *V* is always larger than kappa. With 90% agreement and a 90/10 split in the marginals, however, *V* (.52) is only slightly larger than kappa (.44), as seen in Panel A. Thus, *V,* like kappa, is sensitive to skewed marginals. As with kappa, this sensitivity is even greater at lower levels of interrater agreement (see Panel B). In sum, *V* is only slightly superior to kappa as a metric of interrater agreement for very common or uncommon behaviors. Its performance has not yet been evaluated in Monte Carlo simulations to our knowledge.

Holley and Guilford's G. Holley and Guilford's *G* (1964) is like kappa, except in the manner in which chance agreement is calculated. In contrast to the kappa, in which chance agreement varies with the marginal rates of behaviors, *G* assumes a fixed rate of chance agreement. A generalized form of the equation is given by Gwet (2008):

$$G = (P_o - P_{e|G}) / (1 - P_{e|G}),$$

with $P_{e|G}$ the *G* model expected or chance agreement, given by:

$$P_{e|G} = 1 / q,$$

with *q* equal to the number of response categories. In present/absent comparisons, chance agreement is always .50. Thus, *G* has zero sensitivity to skewed marginals, resulting for example in a value of .80 with 90% agreement and .60 with 80% agreement (Figure 1). According to Gwet's (2008) Monte Carlo simulations, *G* is less biased than kappa.

AC1. Gwet (2002) recently developed the *AC1* statistic as an alternative to kappa that is less sensitive to the skewed marginal problem of kappa, and is given by:

$$AC1 = (P_o - P_{e|AC1}) / (1 - P_{e|AC1}),$$

with *P*o identical to kappa, and with $P_{e|AC1}$ the *AC1* model expected or chance agreement:

$$P_{e|AC1} = 2 * P_+ * (1 - P_+),$$

where P_+ = [(e + g) / 2] / i, and with reference to Table 4. As seen in Figure 1, *AC1* is identical to kappa, *V*, and *G* with balanced marginals. Greater differences between *AC1* and the other metrics emerge at greater splits. Notably, *AC1* assumes somewhat *lower* chance agreement with skewed marginals, in contrast to kappa's opposite assumption. To illustrate, at a 90/10 split and 90% agreement (Figure 1, Panel A), *AC1* is .88 (compared with a kappa of .44), whereas it is .80 with a 50/50 split. Thus, *AC1* does not over-penalize one for skewed marginals. Gwet's (2008) Monte Carlo simulation data suggest that *AC1* produces significantly less biased estimates of interrater agreement than does kappa and slightly outperforms *G*, as well.

Summary and Recommendations. The categorical interrater agreement statistics we have presented produce comparable results when behaviors are neither very frequent nor infrequent—any metric will do in such situations. However, behavioral observations are frequently skewed, and metrics other than Cohen's kappa have been shown to be superior. *G* and *AC1* stand out from kappa and Van Eerdewegh's *V* in this regard, and *AC1* produced less biased estimates of interrater agreement than *G* in Gwet's (2008) Monte Carlo simulations. Thus, we tentatively recommend the *AC1* as the preferred metric. Caution is warranted however as it has not been used widely and has only been evaluated in a single Monte Carlo analysis. Additional study is warranted. Moreover, there are no established rules of thumb for what constitutes, for example, poor and good *AC1*s. However, we believe that it would be reasonable to apply the longstanding criteria for judging kappa (i.e., .60 to .74 is good and .75 and above is excellent; Cicchetti, 1994), with no allowances made for distributional characteristics.

Ordinal, Interval, and Ratio Observations

Intraclass correlation (ICC). The use of the *ICC* for interrater agreement in psychology was popularized by Shrout and Fleiss (1979) and is widely applied to ordinal, interval, and ratio

scaled observations – although technically it assumes interval or ratio data. In simple terms, the *ICC* parses variation in observers' ratings into (a) variance due to differences among the subjects being observed, and (b) variance due to the observers. Interrater agreement, hence the *ICC*, increases to the extent that between subject variance is greater than between observer variance, couched in familiar ANOVA terminology. The *ICC* takes into account both disagreements in the rank ordering of subjects, as well as the means and the variance. To illustrate, if Coder X rates subjects A, B, and C as exhibiting a mean anger intensity of 1, 2, and 3, and Coder Y rates them as 3, 4, and 5, respectively, the *ICC* will punish the disagreement in means (2 vs. 4, respectively), yielding an *ICC* of zero, despite perfect agreement in the rank ordering of the subjects. This example also clearly shows the inadequacy of the Pearson and Spearman correlations for judging interrater agreement, as they are 1.00 despite no absolute agreement in the behavior being rated.

There are several different versions of the *ICC*, raising questions about which to use. Each is estimated in the context of ANOVA, in which variance is segmented into different parcels, such as between subjects (i.e., variation among the people being rated) and within subjects (i.e., variation among the raters). Each easily generalizes to more than two raters.

The most useful *ICC*s for assessing interrater agreement treat coders as a "random effect," meaning that the set of coders used in a given study has been randomly selected from a larger population of coders (Fleiss & Shrout, 1979). It is rarely the case that the particular set of coders who assist us in our research are the only coders of interest and whose ratings in future studies we would like to generalize to; such would be a case for fixed effects analysis.

Whitehurst (1984) suggests the use of a one-way random effects ANOVA, given as:

$$ICC = [MS_B - MS_W] / [MS_B + (k - 1) * MS_W],$$

where MS_B is mean square between subjects, MS_W is mean square within subjects, and k is the

number of judges.

Other writers (e.g., Bakeman & Quera, 2010) suggest the use of one type of 2-way random effects ANOVA; see McGraw and Wong (1996) for others. The primary difference from the one-way approach is that *MS*w is subdivided into its components, MS_E (mean square error) and MS_O (mean square observer). Furthermore, there are two different versions of the 2-way random effects *ICC*. The first is called the "relative consistency" *ICC* and is given by:

$$ICC_{rel} = [MS_B - MS_E] / [MS_B + (k - 1) * MS_E].$$

The second version is called the "absolute agreement" *ICC* and is given by:

$$ICC_{abs} = [MS_B - MS_E] / [(MS_B + (k - 1) * MS_E) + k/n * (MS_O - MS_E)],$$

where *n* is the number of subjects. With the two-way random effects approach, we recommend the absolute agreement *ICC*, the more stringent of the two, in that it reflects more than just whether coders provide similar rank-ordering of the behaviors being rated (i.e., relative agreement), but the degree to which the coders are interchangeable—the highest proof of agreement. The one-way approach is similarly stringent.

Cicchetti's (1994) review suggests the same rules of thumb for interpreting *ICC*s as for categorical metrics (e.g., kappa). However, we recommend a higher criterion: acceptable *ICC*s should exceed .7, and .8 and above is very good. In our experience, *ICC*s below .7 often result from multi-point discrepancies between coders (which suggests the need for more training) or from skewed distributions (which suggests the need to use a different statistic).

Unfortunately, the *ICC* suffers a problem similar to that of Cohen's kappa. The *ICC* is compromised by skewed distributions. As pointed out by Whitehurst (1984), the *ICC* assumes a normal underlying distribution of the trait being measured, with deviations from normality due to rater error. However, many variables of interest in social psychology can be expected to be

skewed. To return to the example of anger intensity ratings, unless an experimental manipulation is unusually powerful, most subjects can be expected to exhibit lower levels of anger, with fewer and fewer subjects showing higher levels of anger – they will likely be positively skewed. Accordingly, the *ICC* is not always the best choice.

The *ICC* can be calculated in common statistical packages (e.g., SPSS), as well as in the "irr" package in the free statistics program, R (cran.r-project.org/web/packages/irr/irr.pdf; R Development Core Team, 2005).

Finn's r. Finn's *r* (Whitehurst, 1984) is an alternative to the *ICC* that is less sensitive to skewed distributions. Also, whereas the *ICC* assumes interval or ratio scaled data, Finn's *r* assumes ordinal structure. This, too, is a positive feature in social psychological research in which observational ratings are often made on single scales that may have only 3, 5, or 7 points, and in which even intervals between scale points cannot be assumed to be even (failing to satisfy criteria for interval scale measurement) and/or do not have a meaningful 0 point (failing to satisfy criteria for ratio scale measurement). Unless such ratings are subsequently averaged (e.g., across multiple experimental periods, similar to items on a questionnaire), ordinal statistical models may be the best fit. Finn's *r* is given by:

$$r_f = (S_c^2 - S_0^2) / S_c^2,$$

where S_c^2 is the expected within-subjects variance when the ratings are assigned randomly and S_0^2 is the MS_W from a one-way random effects ANOVA with independent ratings of each subject as the within subjects variance. $S_c^2$is given by:

$$S_c^2 = (k^2 - 1) / 12,$$

where *k* is the number of ordinal scale categories.

As a rule of thumb, we suggest that Finn's *r* should be above .7 to be considered

acceptable. In our experience, however, Finn's *r* appears impracticably inflated with a greater number of scale categories, including when allowing half-points (i.e., 1, 1.5, … 6.5, 7).

Finn's *r* can be calculated with the above formulas from quantities in the *ICC* output of common statistical programs or with the "irr" package in R (see *ICC* section above).

Weighted Kappa. Weighted kappa, described in the prior section, can be an alternative to the *ICC* or Finn's *r* for ordinal data. As pointed out by Bakeman and Quera (2011), weighted kappa is more easily defensible for ordinal than for nominal data, because the weights assigned to disagreements are less arbitrary in the former case. Disagreements that are further apart on an ordinal rating scale should be penalized more heavily than those that are closer. For example, if aggression in a peer competition task is coded as absent, low, and high, disagreements between ratings of absent vs. high are more serious than absent vs. low or low vs. high disagreements. Linear weights are the most common in such applications. One-point disagreements are assigned a weight of 1, 2-point disagreements a weight of 2, and so on. Quadratic weights, the square of linear weights, are also possible, penalizing far apart differences even more heavily.

Summary and recommendations. The *ICC* is the only choice for truly continuous data. Finn's *r* is a solid alternative to the *ICC* when skewed distributions are a concern, with the proviso that it appears to be inflated when the number of scale points is high. Finn's *r* and weighted kappa are each viable choices for ordinal data.

Interrater Agreement for Sequences

Bakeman et al. (1997) point out that even reasonable levels of behavior-by-behavior interrater agreement does not guarantee that *event sequences* computed from these behaviors are in close agreement. Accordingly, the authors recommend a two-stage process in which interrater agreement is first computed at the level of the behavior, establishing local agreement. Next, the

sequences of interest are computed, using one of the metrics of sequential association (e.g., Yule's Q for categorical data, and the lagged cross-correlation for continuous data). Each subject or dyad has such a value for each sequence of interest. These sequential association metrics are then compared for interrater agreement, using the *ICC*. This approach is very stringent and has not often been used (e.g., Martinez & Forgatch, 2001). Nonetheless, the simulation data of Bakeman et al. (1997) suggest that there can be significant degradation in the interrater agreement of event sequences, compared with local interrater agreement. The two-stage process offers protection against this concern. Bakeman, Quera, and their colleagues offer software for determining event sequence interrater agreement (ELign; Quera, Bakeman, & Gnisci, 2007), which can be downloaded from www2.gsu.edu/~psyrab/BakemanPrograms.htm.

Reliability Across Observations, Contexts, and Time

Generalizability Theory (Gleser, Nanda, & Rajaratnam, 1972) is an extension of the statistical foundations undergirding the ICC. That is, the ICC is based on components of variance due to coders, targets of observation and their interaction. Generalizability theory elegantly partitions variance due to multiple instances within a facet (multiple coders rating the same video) or multiple sources of variance (multiple coders and multiple observations). In the simplest use (multiple coders), Cronbach's alpha can be calculated for an event-coded system, with frequency counts for codes standing in for the "score" on that "test item" and coders standing in for test takers. As an example of using Generalizability Theory for multiple sources of variance, Wieder and Weiss (1981) partitioned variance due to (a) one, two, and four samples and (b) coders in (c) both audio and video conditions. The behavioral samples were collected three weeks apart. For both audio and video samples, most of the variance was accounted for by the "true variance" components (across people and across behavioral samples) and little by the

error sources (coders, first vs. second samples, coder x people, coder x behavioral sample, or coder x people x sample).

How Much Time Is Necessary to Achieve Acceptable Reliability?

As noted above, investigators often use “reliability” and “interrater agreement” interchangeably. Yet, interrater agreement is but one component of stability of measurement (e.g., Hops, Davis, & Longoria, 1995; Kelly, 1977; Mitchell, 1979; Suen, 1988). It is also affected by the stability of the behaviors observed, which is highly dependent on the length of observation. By treating observation intervals as test items, Waters (1978) was able to use conventional psychometric statistics for reliability to determine how long one would have to observe to achieve a set level of reliability (i.e., stable results). In the psychometric theory of test reliability (Cronbach, 1951), test reliability can be assessed via the coefficient of correlation between scores on comparable halves of the test (Ghiselli, Campbell, & Zedeck, 1981). When applying similar principles to observational data, Waters suggested that each 30-second sampling interval be considered a test item which is passed or failed (i.e., the target behavior occurs or does not occur). Interval based coding systems would already be in a form for such statistics. Event-based coding systems can be converted into an interval-based system by using 30-second windows for the events coded. Time intervals can then sorted into odd (1st, 3rd, . . . k) and even (2nd,4th, . . . k–1) groups. The correlation between the odd and even group is the split-half internal consistency reliability for observed variable of interest. (Step-by-step instructions for calculation can be found in the appendix of Heyman et al., 2001.) Heyman et al. (2001) found that in couples conflict observations, 10 to 15 minutes (the most typically used length, established through convention) was sufficient for stable estimates of most codes.

Validity

An important question related to behavioral observation is whether the variables generated are valid measures of the behaviors of interest. Unfortunately, there has been a tendency for researchers to assume that the variables generated from behavioral observation are somehow "more objective, less biased, or inherently superior" (p. 298; Jacobson, 1985) than other measures (such as self-report questionnaires), and this may have limited the examination of the validity of observational measures. Of course, whether or not a measure has more desirable properties than another measure is an empirical rather than philosophical question, and thus cannot be addressed unless data are collected.

The type of validity that is most often cited is face validity. Because behaviors are labeled, and often the labels are relatively straightforward, their validity seems self-evident (thus leading to comments like those of Jacobson). To increase confidence in these variables, however, more information is required. Probably the most important type of validity is construct validity (i.e., whether a tool truly measures what it is intended to measure). This is established via (a) convergent validity, or whether the observed variables (behavior or behavioral pattern) are associated with measures of the same construct that were collected by other means (e.g., "global" self-report, in person interview, diaries or reports of the past 24 hours) and (b) discriminant validity, or whether the observed variables are not associated with measures of different constructs. Predictive validity (whether a tool is related to future outcomes in a hypothesized manner) is especially important in studies that observe behavior to predict outcomes longitudinally (e.g., if roommate conflict early in the year predicts GPA). Finally, discriminative validity (whether a tool can distinguish among groups that are hypothesized to differ) can be used both as a substantive test (e.g., do conflictual and nonconflictual roommate dyads differ on a measure of observed problem resolution?) and as a manipulation check (e.g., does the measure

of obnoxious behavior differ in high and low confederates annoying conditions?)

Analyzing Behavioral Observation Data

When analyzing behavioral data, one must consider both *how* the behavior is measured and *how often* it is measured. In terms of how, for example, behavior can be measured continuously, such as by having observers record impressions of how anxious a participant appeared during an interaction using a 1 (not at all) to 7 (very much scale). Behavior can be measured through a simple count; for example, by having observers log how often a participant blinked during an interaction. The relative nature of one behavior versus others can be measured; for example, the percentage of household labor completed by one partner in a romantic relationship is recorded relative with the other. And lastly, behavior can be measured dichotomously; for example, whether participants wore a condom during sex.

In addition to how behaviors are measured, it is also important to consider how often they are measured. In some cases, each behavior is only measured once for each participant so that data can be analyzed using traditional statistical methods such as regression or ANOVA for continuous outcomes, Poisson regression for count data, or logistic regression for dichotomous outcomes. However, in other cases, behavioral measures are collected several times using a repeated measures design, or they are measured on several occasions over time. For example, during a 15-minute interaction, participants' behaviors may be recorded once per minute, for a total of 15 recordings per participant. In a daily diary study, participants might report on whether they had a fight that day with their romantic partner, for 15 consecutive days. In both of these examples, the data are multilevel because the behaviors of each participant are measured several times, and so time points (or repeated measures) are nested within participants. How participants behaved at one time or repeated measure is likely correlated with how they behaved at another

time or repeated measure, and so the nonindependence in behaviors needs to be adjusted for (Kenny & Kashy, this volume, ch X).

There are several different analytical methods one might employ when analyzing multilevel behavioral data. One strategy that is optimal for many different types of outcomes—linear, count, and dichotomous—is General Estimating Equations (GEE; Liang & Zeger, 1986; Zeger & Liang, 1986). The GEE algorithm is available in most statistics programs (SPSS, SAS, STATA), and Ballinger (2004) provides an excellent description of how to analyze data using GEE. When one is interested in modeling patterns of change over time with continuous measures of behaviors, growth curve models can be estimated using standard multilevel modeling programs, such as the MIXED procedure in SPSS (Proc Mixed in SAS).

As discussed in the section on behavioral observations with experimental manipulations, in some cases, behaviors are measured within dyadic contexts, such as during interactions between romantic partners or between two newly-acquainted partners. When both partners provide behavioral data, their behaviors are likely nonindependent (e.g., how one romantic partner behaves is likely correlated with how her partner behaves within the interaction. In this handbook, Kashy provides an overview of how to analyze dyadic data (see also Kenny & Kashy, 2011; Kenny, Kashy, and Cook, 2006). The same basic principles described in these papers apply to analyzing behavioral data that are dyadic in nature.

Sequential Analysis

Although what people do when interacting is important, how interactions unfold across time is possibly more important. With many phenomena, from the courting behavior of birds to the escalation of human conflict, the patterning of behavior is critical—"a defining characteristic of interaction is that it unfolds in time" (Bakeman & Gottman, 1997, p. 1). Furthermore,

Gottman and Roy (1990, p. 1) contend that: "the dimension of time is so central to conceptualizing social interaction that its use will lead us to think of interaction itself as temporal form."

How did Gottman and colleagues come to conclude that sequence is a central (if not *the* central) issue in understanding behavior? First, Gottman and Roy (1990) discuss several research instances—family management, couples interaction, and schoolchildren's peer interactions—in which base rate analyses show no difference between functional and dysfunctional groups, but analyses of sequence show strong differences between groups. Second, sequential analyses sometimes reveal unexpected patterns and are thus a theory generating, as well as a theory testing tool.

Unidirectional dependence. Most studies that have used sequential analysis have tested if one person's behavior follows the other's behavior at a rate higher than chance. For example, does one roommate's blame follow the other's blame more than what would be expected by chance? This is a one-way, or "unidirectional," test of linkage between the two behaviors.

There are two forms of significant linkage between behaviors. First, compared with chance, the antecedent behavior can increase the likelihood of the consequent behavior. This is an escalation effect. Second, the antecedent behavior can decrease the likelihood of the consequent behavior. This is a suppression effect.

Bidirectional dependence. Bidirectional dependence simultaneously tests if A results in B and if B results in A. For example, we could simultaneously test if roommate A's blame follows roommate B's and if B's blame follows A's blame (i.e., reciprocity). The same logic for escalation and suppression effects applies to bidirectional dependence tests. Wampold (1989) provides a formula for conducting a bidirectional test.

Although one could perform two unidirectional tests, the bidirectional test is superior for three reasons (Wamboldt & Margolin, 1982). First, if one is interested in reciprocity by both partners, the bidirectional test is more appropriate. Second, because unidirectional tests are not independent, multiple tests result in either inflation of the alpha level, or a decrease in power due to the use of the Bonferroni inequality. Third, it is possible for the bidirectional test to be significant, even when each of the unidirectional tests are not.

Dominance. Often researchers are interested in who is the more dominant person in an interaction. Gottman and Ringland (1981, p. 395) defined dominance as an "asymmetry in predictability; that is, if B's behavior is more predictable from A's past [behavior] than conversely, A is said to be dominant." Thus, dominance indicates who is leading the dance. (Note that the label is referring to statistical dominance, which is not necessarily the same as perceived dominance or behavior that might be labeled as domineering.) There are two forms of dominance. In parallel dominance, the same two behaviors are considered. For example, is a student’s hostility more predictable following the teacher’s hostility than the other way around?

Structure of data. To analyze data sequentially, three conditions must be met (Bakeman & Gottman, 1986). First, the temporal sequence must be preserved. Thus, tallies of frequencies (i.e., base rates) are not sufficient; the coding must reflect the order in which the behaviors were performed. Second, codes must be mutually exclusive (i.e., only one code per event). Third, the codes must be exhaustive (i.e., there is a code for each behavior). To construct the matrices needed to test such patterning, it is useful to think of a moving window that can slide over the data stream (only data within the parentheses are visible). Working with a window the size of two events, the analysis would proceed until all the pairs are accounted for. A *transition matrix* (see Table 5) contains the tally of the pairs revealed by the moving window. If the events are not

contiguous, a transition matrix for a specified lag can be computed. The rows specify the lag 0 (i.e., present) behaviors of a wife, and the columns specify the lag 1 (i.e., immediate past) behaviors of her husband. The frequencies represent the number of times that each lag 0 wife behavior is preceded by a husband behavior at lag 1. The example data suggest that husband behavior tends to be reciprocated with like behavior of the wife (e.g., positives are mostly met with positives).

Once the transition matrix is formed, the conditional probability (i.e., the probability of a behavior being emitted, given a particular antecedent behavior) can be computed. If a conditional probability is, say, 0.75, does that constitute an important pattern? This is not known until we know if the conditional probability exceeds chance.

Thus, the null hypothesis in sequential analysis states that the behaviors are randomly ordered and that any apparent patterns are due to chance. A *z*-score—derived by Sacket (1979) and later modified by Allison and Liker (1982) and Wampold and Margolin (1982) —can be computed to test for the deviation from chance. However, despite their widespread use, *z*-scores have a major Achilles heel—they are influenced by the length (total number of transitions in the interaction) and by the base rates of the two behaviors under examination. Thus, the same degree of contingency will produce different *z*-scores across different dyads due to these factors. Non-parametric statistics such as Yule's Q (Bakeman & Quera, 2011) and Wampold Kappa (Wampold, 1989) have been offered and we advise their use. The sequential variables can then be used as scores in calculating test statistics (e.g., correlations, structural equation modeling, multi-level modeling).

Loglinear Approach to Sequential Analysis

Loglinear methods (Bakeman & Quera, 1995) provide a flexible alternative approach to

sequential analysis. The loglinear approach begins with the multidimensional contingency table consisting of frequencies of given behaviors and compares the fit of the observed data to patterns that would result given (the lack of) researcher specified patterns of association. The simplest version is given by the two-way table, an example of which is found in Table 5.

Traditional sequential approaches would model each of the contingencies from Table 5 individually, for example, forming Allison and Liker *z*-scores representing positive→positive, neutral→neutral, and negative→negative contingencies. Such contingencies can be estimated in the loglinear context as well, although via the likelihood ratio chi-square (G^2). However, the loglinear approach is more flexible in that all three contingencies can be tested at once, similar to an omnibus ANOVA testing differences among three groups' means in a single test, thus protecting against Type-I error (Bakeman & Quera, 1995). Loglinear analysis follows the traditional rationale of the chi-square test of independence in which the observed frequencies are compared with expected frequencies that would be obtained if there were no association (a "no two-way interaction model," in loglinear terms). A significant G^2 indicates the presence of a significant lagged association.

The above is a simple example of a loglinear approach to sequential analysis. However, the flexibility of the loglinear approach is that it may be generalized beyond the two-way case to accommodate higher order interactions (e.g., three-way, four-way). For example, one might hypothesize that lag 1 negative husband behavior is less likely reciprocated by the wife at lag 0 if the husband was positive or neutral at lag 2. These frequencies would be represented and tested in a three-way contingency table: lag 0 wife behavior × lag 1 husband behavior × lag 2 husband behavior.

Examples with sophisticated applications of the loglinear approach are found in the study

of in attorney-witness exchanges in the courtroom (Gnisci & Bakeman, 2007) and couples interaction (Notarius et al., 1989). Loglinear analysis can be carried out in standard statistical programs such as SPSS. Bakeman also offers a stand-alone program called ILOG (see Bakeman & Robinson, 1994) at his Internet site: www2.gsu.edu/~psyrab/BakemanPrograms.htm. Furthermore, loglinear analysis can be performed in Mplus (Muthén & Muthén, 2010), as described in the "Multilevel Loglinear Analysis" section below.

Dimensional Analyses of Behavior Sequences

The term "sequential analysis" is usually applied to patterns among categorical observations over the course of an interaction. In many instances, however, researchers are interested in patterns among dimensionally measured behaviors, such as whether the intensity of one person's behavior depends on the intensity of another person's prior behavior. Some examples of this are found in studies of the synchrony of parent-child and adult-adult interaction (e.g., Julien, Brault, Chartrand, & Bégin, 2000; Feldman, 2007; Dowdney & Pickles, 1991; Warner, 1992) and the coordination within and between people, of emotional behavior, emotion experience, and physiology (e.g., Butler, 2011; Guastello, Pincus, & Gunderson, 2006; Levenson & Gottman, 1983; Mauss, Levenson, McCarter, Wilhelm, & Gross, 2005). Such processes usually require a different set of statistics than is used in traditional sequential analysis.

Broadly, dimensional approaches to behavior sequences are usually aimed at establishing the influence of one person's behavior on another person's behavior, the mutual coordination of behaviors in dyads, or sometimes the uninfluenced aspects of social interactions. They make use of what is referred to as *time series* data. A behavioral time series consists of observations made repeatedly at regular intervals of time, such as the intensity of positive emotion rated for each consecutive five-second interval over the course of an interaction. Patterns in observational time

series data can be studied in either the time domain or the frequency domain. The time domain approach is far more common and will thus be emphasized here. Readers interested in the frequency domain approach are referred to Warner (1992) and Richardson, ch x.

Time domain approach. Analyses in the time domain are cast in the familiar terms of correlation and regression, with the primary difference being that the correlations are within-subject or within-dyad. The researcher looks for evidence that present behaviors are correlated between interactive partners and/or that present behaviors are correlated with a person's own past behavior or the past behavior of the partner. The prevailing statistical techniques are cross-correlation and time series regression, although see the "Dynamical Systems Modeling" section below for an alternative approach.

Cross-correlation analysis simply looks for the correlations between one person's behavior in the present (lag 0) with the behavior of another person at various lags (i.e., points in the past). It is up to the investigator to determine which of the various possible cross-correlations s/he is interested in. There may, for example, be substantive reasons for a focus on relatively short or long-term effects. Extensive data preparation is required prior to cross-correlation analysis. Each time series must be "prewhitened," which means that any cycles and trends over time must be removed.

Simply establishing the extent of cross-correlation between behaviors may be of substantive interest. However researchers are often interested in modeling differences in the degree of cross-correlation *among dyads* in relation to other variables. In these cases, the cross-correlation calculated within each dyad with time series methods is used as a variable in other analyses (e.g., Pearson correlation). An example of this is found in Feldman's work in which greater mother-infant synchrony (i.e., greater cross-correlations of mother and infant

behavior) was associated with better child outcomes (e.g., Feldman, 2007).

Time series regression, takes a similar approach to cross-correlational analysis, with two key differences. First, autoregressive effects (i.e., the internal predictability of a person's behavior across time) are part of the model estimation rather than a prior step; such effects are removed when prewhitening variables prior to cross-correlational analysis. Second, because lagged terms are simultaneously entered, each lagged term represents a *unique* association of past and present behavior, controlling for all other lagged effects in the model.

Warner (1992) describes an approach to time series regression for situations in which the researcher wishes to model both the internal (i.e., autoregressive) and social determinants (i.e., partner effects) in behavior times series. The internal and social determinant estimates (R^2) within each dyad may be of substantive interest. However, frequently, investigators are interested in differences in these parameters from dyad to dyad. In such cases, the R^2, like the cross-correlation, can be treated as variables for subsequent analysis. We have used this analytic method to model the impact of child behavior on maternal emotion and how the degree of child influence and the degree of autocorrelation predict maternal discipline practices (Lorber & Slep, 2005).

Time series regression and cross-correlation are available in SPSS Trends (which can be purchased as an add-on) and also in the freely downloadable program, R (R Development Core Team, 2005). R offers a much greater array of time series analytic models and has the added advantage of several automated model selection packages for the prewhitening of data.

Recent Developments in Analyzing Observational Data

In this section, we briefly describe recent analytic developments for observational data that will likely be of interest to many social-psychological researchers. The list is certainly not

comprehensive. Instead, the featured analytic models were selected with an eye toward relevance to social-psychological applications and feasibility of implementation (e.g., availability of computer programs).

Dynamical Systems Modeling. In the mid 1990s, Gottman, Murray, and their colleagues developed a set of nonlinear difference equations—a "dynamical system"—to model change over time in couples behavior via (e.g., Cook et al., 1995). These methods were probably opaque to many researchers and were not implemented in common software packages; they appear to have been used primarily by their progenitors (e.g., Gottman et al., 2003; Gottman, Ryan, Swanson, & Swanson, 2005). However, recent work by Hamaker and colleagues has set the stage for more widespread usage (Hamaker, 2009; Hamaker, Zhang, & Van red Mass, 2009; Madhyastha, Hamaker, & Gottman, 2011).

Briefly, this collection of techniques analyzes dimensional time series data from dyads. The techniques are designed to capture uninfluenced steady states, as well as multiple types of nonlinear influence from one person to another. Uninfluenced steady states refer to what each person "brings to the table," for example, one's overarching emotional style. To illustrate two of the many possibilities for influence: (a) negative behavior in one spouse might have a linear association with the degree of subsequent partner negativity, with rises and falls in one person's behaviors predicting similar rises and falls in the other's behavior, and (b) there could be thresholds above and below which the relation of spousal negativity and subsequent negativity in the partner changes (e.g., a person may "ignore" low level partner negativity, respond in kind to moderate partner negativity, and withdraw from high partner negativity). The innovation of Hamaker et al. (2009) was the realization that Gottman and Murray's models were special cases of the previously established threshold autoregressive model. The advantage is that there are

pre-established influence functions that can be evaluated, methods of parameter estimation, and statistical criteria for selecting from among the different influence models (via the common BIC statistic).

Research using these tools is in its infancy. For example, Madhyastha et al. (2011) showed that many couples do not exhibit reliable interpartner influence (i.e., uninfluenced steady states were very powerful) and that partners from different couples differ in how they are influence each other. Given the availability of the "dyad" statistical package for R (Madhyastha & Hamaker, 2009; R Development Core Team, 2005), a free and very powerful statistical program, there is great untapped potential for the application of nonlinear dynamical modeling in the context of threshold autoregressive models. Such models would be of interest in any social psychological research in which social influence in dimensionally rated behavior dyads might be expected to be nonlinear, and/or in which the estimation of what each dyad member contributes to social interaction, independent of her/his partner's behavior, is of interest.

Multilevel Survival Analysis. Stoolmiller and Snyder (2006; Snyder, Stoolmiller, Wilson, & Yamamoto, 2003) offer a novel approach to characterizing the course of an interaction, utilizing a variant of survival analysis for repeated events based on prior work (Gardner & Griffin, 1989; Griffin & Gardner, 1989). In traditional survival analysis, time to a single event is modeled as the function of predictors or covariates. For example, if one were interested in gender differences in longevity, time to death would be modeled as a function of gender. In contrast, the events of interest in behavioral observation are most often free to repeat. Thus, in the present context, survival analysis is adapted to model repeated events within dyads. The "hazard rate" – time between displays of a behavior – is the focus of these analyses. It can be modeled as a function of static or unchanging covariates (e.g., gender) and time varying

covariates (e.g., behavior of another person and experimental manipulations) via Cox regression, which is one type of survival model.

Hazard rates and their associations with covariates usually vary from dyad to dyad, giving rise to the need to model them in a multilevel framework (Raudenbush & Bryk, 2001; Schoeman & Little, this volume, ch. X). Snyder et al. (2003) provide an example of how multilevel survival analyses can be used to model emotional displays in dyadic interaction. The time between displays of child anger in parent-child interactions (i.e., hazard rate) was modeled as a function of several static and time varying covariates. The authors found that the time between children's anger displays decreased over the course of interactions with their parents the more the parents' insensitive and negative behaviors toward the child accrued, illustrating a time varying covariate effect within dyads. Moreover, children who were rated by their parents as more antisocial (i.e., aggressive and oppositional) had decreased time between anger displays, illustrating a static covariate effect. However, the authors found no evidence that the dynamic link between parenting and child anger was related to parent or teacher-reported antisocial child behavior, illustrating how one might test the association of a static, between dyad covariate (e.g., score on a questionnaire) with a within-dyad dynamic pattern of observed behavior over the course of social interaction.

The multilevel survival approach has, to our knowledge, not yet been employed in social psychology. Nonetheless, Butler (2011) recently pointed out the broad relevance of this approach to what she terms "temporal interpersonal emotion systems" in social interaction, for modeling emotion reciprocity, reactivity, and escalation and de-escalation. The multilevel survival approach has applicability in any setting in which the time between a behavior's occurrence, whether an emotion display or some other behavior, marks a process of theoretical interest.

At present, multilevel survival analyses are available in S-Plus (Insightful Corp., 2001), with S-Plus survival analysis code and SPSS data preparation syntax available from Stoolmiller and Snyder on-line at dx.doi.org/10.1037/1082-989X.11.2.164.supp.

Multilevel Loglinear Analysis. Dagne and Howe (Dagne et al. 2002; Howe et al., 2005) recently developed a multilevel extension of loglinear analysis for observational data (see "Loglinear Approach to Sequential Analysis" section above). This method has multiple advantages over traditional loglinear analyses of behavior observations. To name a few, it has superior handling of the nesting of behavior inherent in many studies of social behavior, where behaviors are often nested within episodes (e.g., experimental conditions) that are further nested within dyads. Multilevel loglinear analysis further deftly handles cases with low rates of target behaviors, a common problem in observational research as estimates of sequential patterns among low rate behaviors have greater measurement error; such cases are weighted to a lesser extent than are higher rate cases. Moreover, it provides an analytic framework to model sequential patterns as a function of other variables.

Howe et al. (2005) use the example of behavior in married couples to illustrate the techniques. Sequences of interest are first estimated within dyads, for example husband reciprocation of wife negativity. Because these patterns occur at different rates in different couples, they are modeled as random effects (e.g., the degree of negative reciprocity is allowed to freely vary among couples). The sample wide average of each random effect (e.g., the overall strength of negative reciprocity) can be compared against zero to test for a significant overall sequential association. The random effects or sequences can then be modeled in relation to other random effects, answering such questions as whether couples who reciprocate one another's positive behaviors at a high rate are less likely to reciprocate negative behaviors. Behavior

sequences or random effects can also be modeled in relation to other consequential variables such as experimental manipulations (e.g., conflict vs. events of the day discussions) and individual or dyad level characteristics (e.g., personality and marital adjustment). Finally, contrasts can also be structured to compare the relative strength of different sequences (e.g., whether men are more likely than women to reciprocate negative behavior).

The multilevel loglinear approach has, to our knowledge, not yet been employed in social psychology. However, it is a very flexible approach with wide applicability to questions of interest of social psychologists who seek to understand behavioral sequences in dyads. Moreover, it is clearly superior to the ordinary loglinear approach to sequential analysis. Howe et al. (2005) offer example syntax for implementing multilevel loglinear analysis in Mplus (Muthén & Muthén, 2010). Moreover, sample Mplus data, input, and output files corresponding to the examples in Dagne et al. (2002) are provided at statmodel.com.

Conclusions and Future Directions

We began this chapter noting that there is nothing so practical as a good theory testing tool. Behavioral observation is a research method centered on the identification of behaviors worth theorizing and enables the testing of theories of behavior. The video and audio records that are created in many behavioral observation studies provide one of the richest sources of information available in the social sciences for the study of social interactions. Because videos (unlike live observations) are archivable, the videos can be used in the future to investigate different hypotheses or as improved methods are developed. Behavioral observation serves as a compliment to a wide variety of self-report methods on behavior, cognition, and affect, as well as an intriguing partner to studies that employ other data collection methods, including biological assays. In short, behavioral observation has great potential for advancing knowledge in social

and personality psychology.

Over the past 50 years, a number of significant advances have occurred in behavioral observation methodology, most notably in terms of the complexity of coding systems, the recording of observations and of codes, and statistical analyses. Each of these advances have been related to advances in computer technology. At present, the field is poised for an explosion in opportunities. With the penetration of smart phones and other digital technology, never has recording been so easy, cheap, and ubiquitous or the opportunities for observation been so plentiful (Mehl & Connor, 2012). Natural sampling methods (like the EAR) will become easier and easier. Further, computerized coding of behavior without the need for human coders (e.g., Black et al., in press; Cohn, Zlochower, Lien, & Kanade, 1999) already exists and will likely increase in its availability and impact in the coming decade.

What has been lacking, however, has been the accumulation of information on the reliability and validity of coding systems, beyond the focus of the field on interrater agreement. Certainly, interrater agreement is vital to the value of a coding system, but it is not the sole issue of interest. At this point, there are a number of general purpose coding systems that have been used by multiple researchers with varied interests over time, and further attention to the basic properties of these systems is needed. Unfortunately, finding funding to do this type of work is difficult and “nuts and bolts” research isn’t flashy. With a renewed focus in this area, and the continued innovation that has been at the core of the method throughout its short history, behavioral observation is well posed to push the field of social and personality psychology forward in the coming years.

Over the past 20 years, the study of behavior in social psychology has rapidly declined, with the majority of studies collecting self-reported measures (e.g., paper and pencil ratings;

Baumeister, Vohns, & Funder, 2007), with some notable exceptions (e.g., behavioral measures of implicit attitudes; see Gawronski and De Houwer, chapter X, this volume, for a review). Scholars may be discouraged from collecting behavioral data in part because of the complexities involved in collecting, coding, and analyzing it. Behavior is also hard to change, and designing an experimental manipulation that alters "actual behavior" may prove daunting for many researchers. For these reasons (and more), researchers may be discouraged from collecting behavioral data. However, behavioral data can provide insight into psychological processes that other dependent measures cannot do alone—it is what put social psychology "on the map" many decades ago and it is at the heart of many of our theories. New scholars should be encouraged to know that there are many basic questions that still remain unanswered in social psychology that can only be answered with behavioral data and that, although not without its challenges, collecting behavioral data is certainly worth the effort.

References

Allison, P. D., & Liker, J. K. (1982). Analyzing sequential categorical data on dyadic interaction: Comment on Gottman. *Psychological Bulletin*, *91*, 393 403.

Altman, I., & Taylor, D. A. (1973). *Social penetration: The development of interpersonal relationships.* New York: Holt, Rinehart, & Winston.

Aron, A., Norman, C., Aron, E., McKenna, C., & Heyman, R. E. (2000). Couples shared anticipation in novel and arousing activities and experienced relationship quality. *Journal of Personality and Social Psychology, 78*, 273-284.

Bakeman, R. & Gottman, J. M. (1997). *Observing interaction: An introduction to sequential analysis* (2nd Ed.). New York: Cambridge University Press.

Bakeman, R., McArthur, D., Quera, V., & Robinson, B. F. (1997). Detecting sequential patterns

and determining their reliability with fallible observers *Psychological Methods, 2,* 357-370.

Bakeman, R., & Quera, V. (1995). Log-linear approaches to lag-sequential analysis when consecutive codes may and cannot repeat. *Psychological Bulletin, 118,* 272- 284.

Bakeman, R., & Quera, V. (2011). *Sequential analysis and observational methods for the behavioral sciences*. New York: Cambridge University Press.

Bakeman, R., & Robinson, B. F. (1994). *Understanding log-linear analysis with ILOG: An interactive approach*. Hillsdale, NJ: Erlbaum.

Ballinger, G. A. (2004). Using generalized estimating equations for longitudinal data analysis. *Organizational Research Methods, 7,* 127-150.

Black, M., Katsamanis, A., Lee, C.C., Lammert, A. C., Baucom, B. R., Christensen, A., Georgiou, P. G., & Narayanan, S. S. (in press). Toward automating a human behavioral coding system for married couples' interactions using speech acoustic features. *Speech Communication.*

Blascovich, J., Mendes, W. B., Hunter, S. B., Lickel, B., & Kowai-Bell, N. (2001). Perceiver threat in social interactions with stigmatized others. *Journal of Personality and Social Psychology 80,* 253-267.

Butler, E. A. (2011). Temporal interpersonal emotion systems: The "TIES" that form relationships. *Personality and Social Psychology Review, 15,* 367-393.

Campos, B., Graesch, A. P., Repetti, R., Bradbury, T., & Ochs, E. (2009). Opportunity for interaction? A naturalistic observation study of dual-earner families after work and school *Journal of Family Psychology, 23,* 798–807. doi:10.1037/a0015824

Christensen, A., & Hazzard, A. (1983). Reactive effects during naturalistic observation of

families. *Behavioral Assessment, 5*, 349-362.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology *Psychological Assessment, 6*, 284-290.

Cohn, J. F., Zlochower, A. J. Lien, J., & Kanade, T. (1999). Automated face analysis by feature point tracking has high concurrent validity with manual FACS coding. *Psychophysiology, 36*, 35-43. DOI: 10.1017/S0048577299971184

Cook, J., Tyson, R., White, J., Rushe, R, Gottman, J., & Murray, J. (1995). Mathematics of marital conflict: Qualitative dynamic mathematical modeling of marital interaction. *Journal of Family Psychology, 9*, 110-130. doi: 10.1037/0893-3200.9.2.110

Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton: Princeton University Press.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

Dadds, M. R., & McHugh, T. A. (1992). Social support and treatment outcome in behavioral family therapy for child conduct problems. *Journal of Consulting and Clinical Psychology, 60*, 252-259.

Dagne, G., Howe, G. W., Brown, C. H., & Muthén, B. (2002). Hierarchical modeling of sequential behavioral data: An empirical Bayesian approach. *Psychological Methods, 7*, 262–280.

Dowdney, L., & Pickles, A. R. (1991). Expression of negative affect with disciplinary

encounters: Is there dyadic reciprocity? *Developmental Psychology, 27*, 606-617. doi: 10.1037/0012-1649.27.4.606

Eid, M. & Diener, E. (Eds.) (2006). *Handbook of multimethod measurement in psychology*. Washington, D.C.: American Psychological Association.

Feldman, R. (2007). Parent-infant synchrony and the construction of shared timing: Physiological precursors, developmental outcomes, and risk conditions. J*ournal of Child Psychology and Psychiatry, 48*, 329–354.

Festinger, L., Riecken, H., & Schacter, S. (1956). *When Prophecy Fails: A Social and Psychological Study of a Modern Group that Predicted the Destruction of the World*. New York: Harper and Row.

Gardner, W., & Griffin, W. A. (1989). Methods for the analysis of parallel streams of continuously recorded social behaviors. *Psychological Bulletin, 105*, 446–455.

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences*. New York: W. H. Freeman.

Gnisci, A., & Bakeman, R. (2007). Sequential accommodation of turn taking and turn length: A study of courtroom interaction. *Journal of Language and Social Psychology, 26*, 234-259.

Goff, P. A., Steele, C. M., & Davies, P. G. (2008). The space between us: Stereotype threat and distance in interracial contexts. *Journal of Personality and Social Psychology, 94*,91-107.

Goodenough, F. L. (1931). *Anger in young children*. Minneapolis, MN: University of Minnesota.

Gottman, J. M. (1978). Nonsequential data analysis techniques in observational research. In G. P. Sackett (Ed.), *Observing behavior: Vol. 2. Data collection and analysis methods* (pp. 45-61). Baltimore: University Park Press.

Gottman, J. M. (1979). *Marital interaction*. Champaign, IL: Research Press.

Gottman, J. M. (1996). *What predicts divorce? The measures*. Hillsdale, NJ: Erlbaum.

Gottman, J. M., Levenson, R. W., Swanson, C., Swanson, K., Tyson, R., & Yoshimoto, D. (2003). Observing gay, lesbian and heterosexual couples' relationships. *Journal of Homosexuality, 45*, 65–91. doi:10.1300/J082v45n01_04

Gottman, J. M., & Krokoff, L. J. (1989). Marital interaction and satisfaction: A longitudinal view. *Journal of Consulting and Clinical Psychology, 57*, 47-52.

Gottman, J., Ryan, K., Swanson, C., Swanson, K. (2005). Proximal change experiments with couples: A methodology for empirically building a science of effective interventions for changing couples' interaction, *Journal of Family Communication, 5*, 163-190.

Gray, H. M., Mendes, W. B., Denny-Brown, C. (2008). An in-group advantage in detecting intergroup anxiety. *Psychological Science, 19,* 1233-1237.

Greenwald, A. G., Nosek, B., & Banaji, M. R. (2003). "Understanding and using the Implicit Association Test: I. An improved scoring algorithm": Correction to Greenwald et al. (2003). *Journal of Personality & Social Psychology, 85*, 41.

Griffin, W. A. (2000). A conceptual and graphical method for converging multisubject behavioral observational data into a single process indicator. *Behavior Research Methods, Instruments, and Computers*, *32*, 120-133.

Griffin, W. A., & Gardner, W. (1989). Analysis of behavioral durations in observational studies of social interaction. *Psychological Bulletin, 106*, 497–502.

Guastello, S. J., Pincus, D., & Gunderson, P. R. (2006). Electrodermal arousal between participants in a conversation: Nonlinear dynamics and linkage effects. *Nonlinear Dynamics, Psychology, and Life Sciences, 10*, 365–399.

Gwet, K. (2002). Kappa statistic is not satisfactory for assessing the extent of agreement between

raters. *Statistical Methods for Inter-rater Reliability,* 1, 1-5.

Gwet, K. (2008). Variance estimation of nominal-scale inter-rater reliability with random selection of raters. *Psychometrika, 73*, 407-430.

Hamaker, E. L. (2009). Determining the number of regimes in threshold autoregressive models by means of information criteria. *Journal of Mathematical Psychology, 53*, 518–529. doi:10.1016/j.jmp.2009.07.006

Hamaker, E. L., Zhang, Z., & Van der Maas, H. L. J. (2009). Using threshold autoregressive models to study dyadic interactions. *Psychometrika*, *74*, 727–745. doi:10.1007/s11336-009-9113-4

Haney, C., Banks, W., & Zimbardo, P. (1973). Interpersonal dynamics in a simulated prison. *International Journal of Criminology and Penology, 1*, 69-97.

Hawes, D. J., & Dawes, M. R. (2006). Assessing parenting practices through parent-report and direct observation during parent-training. *Journal of Child and Family Studies, 15* (5), 555-568.

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

Heyman, R. E. (2001). Observation of couple conflicts: Clinical assessment applications, stubborn truths, and shaky foundations. *Psychological Assessment, 13*, 5-35.

Heyman, R. E. (2004). Rapid Marital Interaction Coding System. In P. K. Kerig & D. H. Baucom (Eds.) *Couple observational coding systems* (pp. 67-94). Mahwah, NJ: Lawrence Erlbaum Associates.

Heyman, R. E., Eddy, J. M., Weiss, R. L., & Vivian, D. (1995). Factor analysis of the Marital Interaction Coding System. *Journal of Family Psychology, 9*, 209-215.

Hops, H., Davis, B., & Longoria, N. (1995). Methodological issues in direct observation: Illustrations with the Living in Familial Environments (LIFE) coding system. *Journal of Clinical Child Psychology, 24*, 193–203.

Holley, W., & Guilford, J. P. (1964). A note on the G-index of agreement. Educational and *Psychological Measurement, 24*, 749–754

Howe, G. W., Dagne, G., & Brown, C. H. (2005). Multilevel methods for modeling observed sequences of family interaction. *Journal of Family Psychology, 19*, 72-85.

Insightful Corp. (2001). *S-Plus 6 for Windows user's guide*. Seattle: Author.

Jacobson, N. S. (1985). The role of observational measures in behavior therapy outcome research. *Behavioral Assessment, 7*, 297-308.

Julien, D., Brault, M., Chartrand, E., & Begin, J. (2000). Immediacy behaviors and synchrony in satisfied and dissatisfied couples. *Canadian Journal of Behavioural Science, 32*, 84–90.

Kenny, D. A., Kashy, D. A. (2011). Dyadic data analysis using multilevel modeling. In J. J. Hox & J. K. Roberts (Eds.), *Handbook for advanced multilevel analysis* (pp. 335-370). New York: Routledge/Taylor & Francis Group.

Kenny, D. A., Kashy, D. A., & Cook, W. L. (2006). *Dyadic data analysis*. New York: Guilford.

Kerig, P. K. & Baucom, D. H. (Eds.) (2004). *Couple observational coding systems*. Mahwah, NJ: Lawrence Erlbaum Associates.

Kerig, P. K. & Lindahl, K. M. (Eds.) (2000). *Family observational coding systems*. Mahwah, NJ: Lawrence Erlbaum Associates.

Kenny, D. A., Mohr, C. D., & Levesque, M. J. (2001). A social relations variance partitioning of dyadic behavior. *Psychological Bulletin, 127,* 128-141.

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika, 73,* 13-22.

Lorber, M. F. (2006). Can minimally trained observers provide valid global ratings? *Journal of Family Psychology, 20*, 335-338.

Lorber, M. F., & Slep, A. M. S. (2005). Mothers' emotion dynamics and their relations with harsh and lax discipline: microsocial time series analyses. *Journal of Clinical Child and Adolescent Psychology, 34*, 559-568.

Lorenz, K. (1970). *Studies in Animal and Human Behaviour. v.1*. (R. Martin, Transl.). Cambridge, Massachusetts: Harvard University Press.

Lorenz, K. (1971). *Studies in Animal and Human Behaviour. v.2*. (R. Martin, Transl.). Cambridge, Massachusetts: Harvard University Press.

Madhyastha, T. M., Hamaker, E. L., & Gottman, J. M. (2011). Investigating spousal influence using moment-to-moment affect data from marital conflict. *Journal of Family Psychology, 25*, 292-300.

Madhyastha, T., & Hamaker, E. (2009). *Dyad*. Retrieved from http://cran.r-project.org/web/packages/dyad/index.html

Margolin, G., Oliver, P., Gordis, E., O'Hearn, H. G., Medina, A. M., Ghosh, C. M., & Morland, L. (1998). The nuts and bolts of behavioral observation of marital and family interaction. *Clinical Child and Family Psychology Review1*, 195–213.

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.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*, 30-46. doi: 10.1037/1082-989X.1.1.30

Mead, M. (1928) *Coming of age in Samoa: A psychological study of primitive youth for Western civilisation*. New York: William Morrow & Co.

Mehl, M. R. (2007). Eavesdropping on health: A naturalistic observation approach for social health research. *Social and Personality Psychology Compass, 1*, 359–380. doi:10.1111/j.1751-9004.2007.00034.x

Mehl, M. R., Gosling, S. D., & Pennebaker, J. W. (2006). Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. *Journal of Personality and Social Psychology, 90*, 862–877.

Mehl, M. R. & Pennebaker, J. W. (2003). The sounds of social life: A psychometric analysis of students' daily social environments and natural conversations. *Journal of Personality and Social Psychology, 84*, 857–870.

Mehl, M. R., Pennebaker, J. W., Crow, M. D., Dabbs, J., & Price, J. H. (2001). The Electronically Activated Recorder (EAR): A device for sampling naturalistic daily activities and conversations. *Behavior Research Methods, Instruments, and Computers, 33*, 517-523.

Mehl, M. R. & Conner, T. S. (Eds.), *Handbook of research methods for studying daily life*. New York, NY: Guilford Press

Mehl, M. R. & Robbins, M. L. (2012). Naturalistic observation sampling: The Electronically Activated Recorder (EAR). In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 176-192). New York, NY: Guilford Press

Mehl, M. R., Vazire, S., Ramírez-Esparza, N., Slatcher, R. B., & Pennebaker, J. (2007). Are women really more talkative than men? *Science, 317*, 82.

Mitchell, S. (1979). Interobserver agreement, reliability, and generalizability of data collected in observational studies. *Psychological Bulletin, 86*, 376–390.

Muthén, L. K., & Muthén, B. O. (2010). *Mplus User's Guide* (6th Ed). Los Angeles, CA: Author.

Notarius, C. I., Benson, P. R., Sloane, D., Vanzetti, N. A.; et al (1989). Exploring the interface

between perception and behavior: An analysis of marital interaction in distressed and nondistressed couples. *Behavioral Assessment, 11*, 39-64.

Patterson, G.R. (1982). *Coercive family process*. Eugene, OR: Castalia.

Patterson, G. R. Reid, J. B., & Dishion, T. J. (1992). *Antisocial boys.* Eugene, OR: Castalia.

Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, B. D. (2005). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology, 89,* 277-293.

Penner, L. A., Dovidio, J. F., Piliavin, J. A., & Schroeder, D. A. (2005). Prosocial behavior: Multilevel perspectives. *Annual Review of Psychology, 56*, 365–392.

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

Reid, J. B. (Ed.) (1978). *A social learning approach, Vol. 2: Observation in home settings*. Eugene, OR: Castalia.

Reid, J. B., Patterson, G. R., & Snyder, J. J. (2002). *Antisocial behavior in children and adolescents: A developmental analysis and the Oregon model for intervention*. Washington, DC: APA Press.

Rusby, J., Estes, A. & Dishion, T.J. (1991). *Interpersonal process code*. Unpublished coding manual, Oregon Social Learning Center, Eugene, OR.

Sackett, G. P. (1979). The lag sequential analysis of contingency and cyclicity in behavioral interaction research. In J. D. Osofsky (Ed.), *Handbook of infant development* (pp. 623 649). New York: Wiley.

Schmaling, K. B., Wamboldt, F., Telford, L., Newman, K. B., Hops, H., & Eddy, J. M. (1996). Interactions of asthmatics and their spouses: A preliminary study of individual

differences. *Journal of Clinical Psychology in Medical Settings, 3*, 211-218.

Shelton, K. K., Frick, P. J., & Wootton, J. M. (1996). Assessment of parenting practices in families of elementary school-aged children. *Journal of Clinical Child Psychology, 25*, 317-329.

Shelton, J. N. & Richeson, J. (2006). Interracial interactions: A relational approach. In M. P. Zanna (Ed.), *Advances in Experimental Social Psychology, Vol. 38*. (pp. 121-181). San Diego: Elsevier Academic Press.

Shrout, P. E. & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420–428. doi:10.1037//0033-2909.86.2.420

Shumway, R. H., & Stoffer, D. S. (2010). *Time series analysis and its applications: With R examples* (3rd ed.). New York, NY: Springer.

Smith, R. H. & Harris, M. J. (2006). Multimethod approaches in social psychology: Between- and within-method replication and multimethod assessment. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 385-400). Washington, D.C.: American Psychological Association. doi: 10.1037/11383-026.

Smoak, N. D, Scott-Sheldon, L. A. J., Johnson, B. T., Carey, M. P. (2006). Sexual risk reduction interventions do not inadvertently increase the overall frequency of sexual behavior: A meta-analysis of 174 studies with 116,735 participants. *Journal of Acquired Immune Deficiency Syndromes*, *41*, 374-384.

Snyder, J., Edwards, P., McGraw, K., Kilgore, K., & Holton, A. (1994). Escalation and reinforcement in mother-child conflict: Social processes associated with the development of physical aggression. *Development and Psychopathology, 6*, 305-321.

Snyder, J., Stoolmiller, M., Wilson, M., & Yamamoto, M. (2003). Child anger regulation,

parental responses to children's anger displays, and early child antisocial behavior. *Social Development, 12*, 335–360.

Spitznagel, E. L., & Helzer, J. E. (1985). A proposed solution to the baserate problem in the kappa statistic. *Archives of General Psychiatry, 42*, 725-728.

Stoolmiller, M., & Snyder, J. (2006). Modeling heterogeneity in social interaction processes using multilevel survival analysis. *Psychological Methods, 11*, 164-177.

Suen, H. K. (1988). Agreement, reliability, accuracy, and validity: Toward a clarification. *Behavioral Assessment, 10*, 343–366.

Tashakkori, A., & Teddlie, C. (2010). *Handbook of mixed methods in social and behavioral research (2nd Ed.)*. Thousand Oaks, CA: Sage.

Thornberry, T., & Brestan-Knight, E. (2011). Analyzing the Utility of Dyadic Parent-Child Interaction Coding System (DPICS) Warm-Up Segments. *Journal of Psychopathology and Behavioral Assessment, 33*, 187–195. doi:10.1007/s10862-011-9229-6

Tickle-Degnen, L., & Rosenthal, R. (1990). The nature of rapport and its nonverbal correlates. *Psychological Inquiry, 1*, 285-293.

Van Baaren, R. B., Janssen, L., Chartrand, T. L., & Dijksterhuis, A. (2009). Where is the love? The social aspects of mimicry. *Philosophical Transactions of the Royal Society B, 1528,* 2381–2389.

Wampold, B. E. (1989). Kappa as a measure of pattern in sequential data. *Quality & Quantity, 23*, 171–187.

Wampold, B. E., & Margolin, G. (1982). Non parametric strategies to test independence of behavioral states in sequential data. *Psychological Bulletin, 92*, 755 765.

Wang, S. W., Repetti, R. L., & Campos, B. (2011). Job stress and family social behavior: The

moderating role of neuroticism. *Journal of Occupational Health Psychology, 16*, 441–456. doi:10.1037/a0025100

Warner, R. M. (1992). Sequential analysis of social interaction: Assessing internal versus social determinants of behavior. *Journal of Personality and Social Psychology, 63*, 51-60.

Waters, E. B. (1978). The reliability and stability of individual differences in infant-mother attachment. *Child Development, 49*, 483–494.

Whaley, S. E., Pinto, A., & Sigman, M. (1999). Characterizing interactions between anxious mothers and their children. *Journal of Consulting and Clinical Psychology, 67*, 826-836.

Wieder, G. B., & Weiss, R. L. (1980). Generalizability theory and the coding of marital interactions. *Journal of Consulting and Clinical Psychology, 48*, 469-477.

Weiss, R. L., & Summers, K. J. (1983). Marital Interaction Coding System III. In E. E. Filsinger (Ed.) *A sourcebook of marriage and family assessment* (pp. 85 115). Beverly Hills: Sage.

Whitehurst, G. J. (1984). Interrater agreement for journal manuscript reviews. *American Psychologist, 39*, 22–28.

Zeger, S. L., & Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous out-comes. *Biometrics, 42*, 121-130.

Table 1.
Rules for Quasi-Naturalistic Family Observation Sessions

1. Everyone in the family must be present.
2. No guests.
3. The family is limited to two rooms.
4. The observers will wait only 10 minutes for all to be present in the two rooms.
5. Telephone: No calls out; briefly answer incoming calls.
6. No TV.
7. No talking to observers while they are coding.
8. Do not discuss anything with the observers that relates to your problems or the procedures you are using to deal with them.

Notes. From Reid, 1978, p. 8

Table 2

Dyadic conflict discussion protocol

1. Setup (prior to 1st interaction)
    a. Check random number list to determine if the topic from Participant 1 (e.g., woman) or Participant 2 (e.g., man) topic is first.
    b. Look at each participant's top areas of conflict (e.g., from Areas of Change Questionnaire). Pick top area of desired change for participant who will initiate first conversation. In case of a tie within a person, use random number sheet to determine order. If both participants pick the same topic, use it for whomever is randomly chosen to go first. Then choose the next highest topic for the second participant's discussion.
2. Instructions for conversations are given separately to participants (i.e., they are in different rooms).
    a. To the participant who will initiate the discussion, begin with "You wrote that you'd like to see [other participant's name] change [conflict topic]…"
    b. To the other partner, begin with "Your partner wrote that s/he'd like to see you change [conflict topic]…"
    c. "We'd like you to have a conversation with [name] about that topic for 10 minutes and try to get somewhere with it. We'd just like to see you discuss this like you typically talk about problems when you are [at home/in your dorm room/etc.]. [pause for questions] OK, we're just about ready. The last thing is to make sure that you know how you will start. Think to yourself about what you would do if you were to bring up [conflict topic] [at home/in your dorm room/etc.]. Do you know how you would start?" [Check to make sure that she have some way to start]
3. Prior to 2nd interaction,
    a. To the participant who will initiate the discussion, begin with "You wrote that you would like to see to see [other participant's name] change [conflict topic]…"
    b. To the other partner, begin with "Your partner wrote that he/she would like to see you change [conflict topic]…"
    c. To both: "…We'd like you to have a conversation with [name] about that topic for 10 minutes and try to get somewhere with it. Like last time, we'd just like to see you discuss this like you do at home/in your dorm room/etc.]."

Table 3

Coding Units

Sampling Unit	**What is Recorded**	**Example**
Event	The occurrence of each behavior of interest.	Noting each time a smile occurs over a 10-minute recording period.
Duration	The length of each behavior of interest (behavior onset and offset times).	Noting the total length of time smiling occurs over a 10-minute recording period.
Interval	The occurrence of each behavior of interest in each consecutive time block/interval.	The presence/absence of similing is noted for each 5-second interval during a 10-minute recording period.
Time	Intermittent observations, typically using duration or interval sampling, and the occurrence (and sometimes the frequency) of behaviors of interest.	Using event, duration, or interval sampling of smiling but only every other minute during a 10-minute recording period.

Table 4

Confusion Matrix for Presence vs. Absence of a Behavior Rated by Two Coders

	Coder 2		
Coder 1	Behavior Present	Behavior Absent	Row (Coder 1) Totals
Behavior Present	**a**	b	a + b = e
Behavior Absent	c	**d**	d + e = f

Column (Coder 2) Totals	a + d = g	b + e = h	a + b + c + d = i

Table 5

Contingency Table (Transition Matrix) of Lagged Effects of Dyadic Behavior

	Lag 1 Partner 2 Behavior		
Lag 0 Partner 1 Behavior	Positive	Neutral	Negative
Positive	20	10	0
Neutral	10	20	10
Negative	0	10	20

71

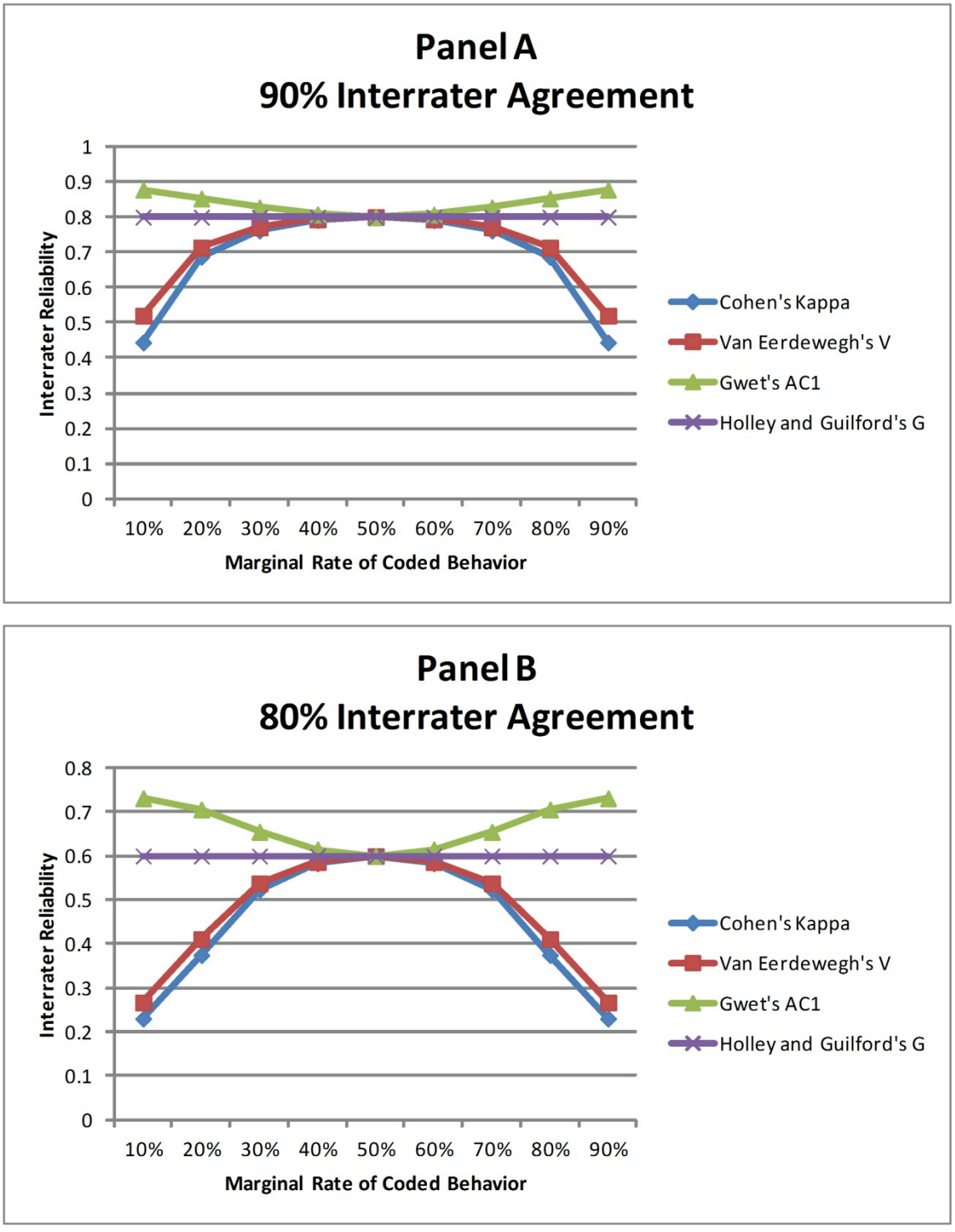

Figure 1. The performance of five interrater agreement statistics across different marginal rates of behavior, and for 90% (Panel A) and 80% (Panel B) raw interrater agreement.