Published in final edited form as:
Behav Ther. 2001 ; 32(1): 107–122.

How Much Observational Data Is Enough? An Empirical Test Using Marital Interaction Coding

Richard E. Heyman, **Bushra R. Chaudhry**, **Dominique Treboux**, **Judith Crowell**, **Chiyoko Lord**, **Dina Vivian**, and **Everett B. Waters**
State University of New York at Stony Brook

Abstract

Using three different samples of couples (clinic, nondistressed community, and engaged), we found that 15 minutes was sufficient to witness enough behavior to make reliable (i.e., internally consistent) estimations of most Rapid Marital Interaction Coding System (Heyman & Vivian, 1993) code frequencies. Ten minutes is sufficient for many codes of interest. The ease in which "how much time is necessary" calculations can be made should entice behavioral investigators from a variety of content areas to publish such figures. By empirically investigating a factor that in most fields becomes reified through convention, investigators can conduct observational research that is both maximally efficient and maximally scientifically defensible.

Direct observation of behaviors of interest is so closely identified with behavioral approaches to assessment and treatment that the editor of *Behavior Therapy* had to warn recently that the journal does not, in fact, *require* observations for manuscripts to be considered appropriate for publication (Beck, 1999). The strength of direct observation is that actual events are being recorded and described, with the description meeting scientific standards for replicability (i.e., interrater agreement: "for properly trained observers to produce identical protocols, given that they observed the same stream of behavior"; Bakeman & Gottman, 1997, pp. 2–3). Direct observation "can provide measures of responses that most subjects cannot accurately describe, such as behavior rates, expressive movements, and fleeting movements, and for events that subjects may be unwilling to report or may distort. . . ." (Hartmann & Wood, 1990, p. 108). Finally, direct observation is particularly adept at investigating the processes by which self-reported distress (e.g., marital dissatisfaction, anxiety) develops and is maintained.

However, observational data is expensive and difficult to collect and code. Investigators since the dawn of this form of research have asked themselves "How long do I need to observe?" Like most methodological questions, the answer is not only much more complex than it seems at first blush, but is best answered with the (somewhat infuriating) reply: "It depends." It depends on whom you're observing, on what you're interested in testing, and on the costs to both the participants and to the experimenter. Who you're observing is important because developmental stages (of individuals or couples, or both) and clinical statuses of participants will affect the types and frequencies of behaviors emitted (e.g., Bradbury, 1998). What you're interested in testing is important because the frequency of the particular behaviors under examination will affect the length of time required. Costs are important because expensive observations should only be as long as they need to be in order to get and code the necessary data. Finally, answering "How long do I need to observe?" also has far-reaching implications on the use of observational measures in clinical practice. For example, if distressed couples'

Address correspondence to Richard E. Heyman, Department of Psychology, State University of New York, Stony Brook, NY 11794-2500; e-mail:richard.heyman@sunysb.edu..

This research was supported by Grants R01MH44935, R01MH57779, and R29MH44665 from the National Institutes of Mental Health.

communication styles can be discerned reliably in 10 to 15 minutes, it is reasonable to recommend that such observations be included in all clinical assessments of couples, regardless of whether they are participating in a clinical research study. However, if 1 hour is required, such a recommendation may put an undue burden on clinicians and couples, tilting the cost-benefit balance away from observation.

Questions of observation length are embedded in larger questions of reliability and validity, which themselves address the larger question of generalizability. This paper will focus specifically on marital observation, although much that we will cover has direct parallels in most areas of clinical psychology.

Generalizability of observed marital interactions was studied by Wieder and Weiss (1980) across the facets of coders, multiple interactions, and couples. They found strong evidence that the variance in marital interaction coding is due to differences among couples, not to the other facets. External validity has been established by the substantial similarity of home and laboratory observations (Gottman, 1979) and by couples' reports that their observed interactions are typical of their home interactions (Foster, Caplan, & Howe, 1997; Margolin, John, & Gleberman, 1988). Validity has also been indicated by (a) the broad replication of findings across laboratories in North America, United Kingdom, Germany, Spain, and Australia; (b) the ability of marital observations to discriminate between distressed and nondistressed couples; (c) the ability of marital observations to predict risk for future marital declines and separation/divorce (see Heyman, in press, for a review of all published reliability and validity findings).

Reliability has received almost no attention, however. Investigators often use reliability and interrater agreement interchangeably. Although interrater agreement can be considered a form of reliability, understanding the stability of coded results requires more than an estimate of agreement. Although this has been discussed before (e.g., Hops, Davis, & Longoria, 1995; Kelly, 1977; Mitchell, 1979; Suen, 1988), it still has not received adequate attention in the behavioral observation literature.[1] In the case of marital observation, Wieder and Weiss's (1980) study demonstrated that results are fairly stable across conflict tasks, and Gottman's (1979, 1980) study demonstrated that results are fairly stable between home and laboratory settings. As noted above, however, stability is dependent on who is being studied, on how frequent the codes of interest are, and on how long the observations are. Because the first two factors are often the independent and dependent variables of the study, the length of observation is typically an invariant, methodological decision made by the investigator.

Investigators, however, have little empirical guidance in making this decision. An accepted paradigm typically evolves and becomes reified, even if it is based more on tradition than science (Kuhn, 1970). The accepted marital interaction paradigm involves a 10- to 15-minute problem-solving interaction task. Given the generalizability and validity studies alluded to above, this paradigm seems to have empirical support. But why 10 to 15 minutes? In contrast, family observational studies (e.g., Patterson, 1982) typically collect at least ten 1-hour home observations.

[1]Mitchell (1979) provides a cogent discussion of the critical distinctions between interobserver agreement and reliability: "[I]t should be repeated that reliability and observer agreement are not the same . . . The differences between agreement and reliability are based on the way the two indices are defined. Reliability coefficients partition the variance of a set of scores into a true score (individual differences) and an error component. The error component may include random fluctuations in the behavior of subjects, inconsistencies in the use of the scale, differences among observers, and so forth. Interobserver agreement [indices], on the other hand, carry no information at all about individual differences among subjects and contain information about only one of the possible sources of error—differences among observers. In other words, a reliability coefficient reflects the relative magnitude of all error with respect to true score variability, whereas an agreement percentage reflects the absolute magnitude of just one kind of error" (p. 378).

The reliability (representativeness) issue has been raised frequently outside of the observational area, especially in the personality and individual differences literature. As noted frequently (e.g., Block, 1977; Epstein, 1983; Mischel, 1968), low correlations in longitudinal and construct validity research can be due to the low quality of the data, particularly the low reliability of many measures. Waters (1978) has made this same point in developmental literature in relation to observational data. Waters examined representativeness by treating observation intervals as test items and using conventional psychometric statistics to examine reliability. When "test items" are actually interval samples, the amount of time necessary to obtain reliable estimates of subjects' typical rates of specific behaviors can be estimated. Waters concluded that in a given observation sample, estimates of reliability are lower for low base rate behaviors and that the amount of time needed to obtain reliable estimates often far exceeds the duration of observation intervals used in both laboratory and naturalistic observational studies. Low reliability attenuates correlations in individual differences studies and reduces statistical power in between-group comparisons. More importantly, unequal reliabilities across behavior categories differentially attenuate results and thus distort patterns of results (i.e., not a simple underestimation of effects, but the wrong pattern of effects, especially when multivariate analyses are used).

Waters (1978) observed infants with their mothers and with strangers. This paper applies similar reliability-estimation methods to answer the following questions: How long must one observe to obtain a stable (i.e., reliable) estimate of the frequency of coded marital behavior? Does the type of sample affect the length of observation needed?

Method

Participants

Sample 1—One-hundred ninety-seven couples (clinic subsample) presented for treatment at the University Marital Therapy Clinic at the State University of New York at Stony Brook as part of a larger study (see Boyle & Vivian, 1996; Cascardi & Vivian, 1995; Ehrensaft & Vivian, 1996; Vivian & Langhinrichsen-Rohling, 1994). An additional 50 happily married couples (community subsample) were recruited (through advertisements in local papers) to serve as a control group for the clinical subsample described above. Couples had to meet the following criteria: (a) neither partner reported any acts of physical aggression during the past year on the Conflict Tactics Scale (Straus, 1979); (b) both partners reported in a clinical interview that husband-to-wife physical aggression had never occurred in the relationship; and (c) both partners scored above 97 on the Dyadic Adjustment Scale (DAS; Spanier, 1976). The subsamples had the following demographic characteristics—married: clinic, 85%, community, 90%; mean length of marriage: clinic, 10.98 years (*SD* = 12.79), community, 13.39 years (*SD* = 12.20); men's mean age: clinic, 37.92 (*SD* = 11.68), community 39.67 (*SD* = 12.39); women's mean age: clinic, 34.70 (*SD* = 11.10), community 38.47 (*SD* = 11.72), full-time employment: clinic men, 71.0%, community men, 74.5%, clinic women, 29.5%, community women, 32.0%; mean family income: clinic, $43,580.70 (*SD* = $29,097.15), community: $50,728.26 (*SD* = $25,314.98). The mean DAS scores were as follows: clinic men, 86.87 (*SD* = 18.39), community men, 118.89 (*SD* = 11.66), clinic women 79.73, (*SD* = 23.26), community women, 122.26 (*SD* = 12.12).

Sample 2—This sample was drawn from a longitudinal study of adult attachment over the course of marriage (Waters & Crowell, 1990). The full sample comprises 157 couples assessed 3 months prior to their wedding dates. Because of technical difficulties with the videotapes, the sample for this study comprises 142 participants. The sample is predominantly white (96%). The mean age for women and men was 23.3 years (*SD* = 1.5) and 24.7 years (*SD* = 2.2), respectively. The majority of couples (79%) were employed full-time; median occupational

status of 6 (e.g., technician, semi-professional, or small business owner) using Hollingshead's (1975) rating. The mean number of years of education was 14.8 years. The mean duration of the relationship was 50 months (*SD* = 25.0). None of the subjects had been married before and they had no children at the time of recruitment. Participants reported high satisfaction in their relationship on a 7-point Likert-type item (identical to Item 31 of the DAS asking participants to indicate their degree of happiness; $X = 4.7$, $SD = 1.8$).

Measures

Rapid Marital Interaction Coding System (RMICS; Heyman & Vivian, 1993)—The RMICS[2] is an observational coding system adapted from the Marital Interaction Coding System–IV (MICS-IV; Heyman, Weiss, & Eddy, 1995). In a study comparing the full 37-code MICS and the 11-code RMICS, the RMICS performed favorably (Heyman, Vivian, Weiss, Hubbard, & Ayerle, 1993). The RMICS was based on a factor analysis of all 1,088 couples coded with the MICS over a 5-year period. The original 37 microbehavioral MICS codes formed four factors—hostility, constructive problem discussion, humor, and responsibility discussion (Heyman, Eddy, Weiss, & Vivian, 1995). The first three factors became RMICS codes, whereas the fourth (responsibility discussion) was incorporated into the broader notion of attributions. We used Holtzworth-Munroe and Jacobson's (1988) distillation of attributions into distress-maintaining and relationship-enhancing attribution codes. Further, we added several codes to make the system exhaustive and content valid. We included two codes added to the original MICS after the factor analysis was conducted—withdrawal and dysphoric affect (Heyman, Weiss, et al., 1995). Two positive codes (self-disclosure and acceptance) were also incorporated from a similar marital coding system, the Kategorien-system fur Partnerschaftliche Interaktion (KPI; Hahlweg et al., 1984).[3]

The speaker turn is the RMICS's basic coding unit. The codes are ordered hierarchically, based on both communication theory and substantial research that demonstrates that negative, followed by positive, followed by neutral, codes are of decreasing importance in understanding marital conflict (see Weiss & Heyman, 1997). If a participant emits more than one code during a speaker turn, she or he receives the code highest on the hierarchy. To deal with long monologues, speaker turns that last more than 30 seconds are interval coded in 30-second segments (i.e., coded as if a new speaker turn occurs every 30 seconds). In declining hierarchical importance, the RMICS comprises psychological abuse (e.g., demeaning statements); distress-maintaining attributions (negative causal explanations); hostility (e.g., angry affect, criticism, combativeness); dysphoric affect (e.g., sad affect); withdrawal (e.g., stonewalling); relationship-enhancing attributions (positive causal explanations); acceptance (e.g., paraphrasing, expressions of caring); self-disclosure (statements that express the speaker's feelings, wishes, or beliefs; acceptance of responsibility); humor (e.g., joking, laughing); constructive problem discussion (e.g., description of the problem, constructive solutions, questions and agreement); other (statements on something other than a personal or relationship topic; e.g., "Is that the camera?").

Four coders were assigned to code videotapes from Sample 1. Videotapes were randomly assigned to each coder, with 41 of the 267 tapes (17%) randomly assigned to two coders to provide interrater reliability checks. Inter-rater reliability (i.e., overall Cohen's for the entire system, averaged across the 41 interrater reliability tapes) was 0.57 (*SD* = 0.16; kappas-by-code were calculated by combining all 41 reliability tapes into one confusion matrix, and then calculating kappa for each code. (This was done to compensate for low base rates of any particular code during a single session.) Kappas-by-code were excellent for a system of this

[2]A copy of the RMICS is available from the authors, and can be found on the first author's Web site: http://www.psy.sunysb.edu/marital.
[3]Although RMICS-based studies are just now being published, it has been used on over 2,000 couple interactions by over a dozen labs. The studies will be published in the ensuing years.

type: distress-maintaining attributions (κ = .69); hostility (κ = .71); dysphoric affect (κ = .89); withdrawal (κ = .51); relationship-enhancing attributions (κ = .67); acceptance (κ = .57); self-disclosure (κ = .70); humor (κ = .79); and constructive problem discussion (κ = .69).

Three coders were assigned to code videotapes from Sample 2. Videotapes were randomly assigned to each coder, with 32 of the 142 tapes (23%) randomly assigned to two coders to provide interrater reliability checks. Inter-rater reliability (i.e., overall Cohen's κ for the entire system, averaged across the 32 interrater reliability tapes) was 0.55 (*SD* = 0.16). Because engaged couples, compared to the full range of married couples, emit a more narrow range of behaviors during conflict interactions, kappa is an extremely conservative measure: Mean percentage agreement was .80 (*SD* = .07). Kappas-by-code were as follows: distress-maintaining attributions (κ = .65); hostility (κ = .67); dysphoric affect (too infrequent to calculate); withdrawal (too infrequent to calculate); relationship-enhancing attributions (κ = .43); acceptance (κ = .47); self-disclosure (κ = .56); humor (κ = .54); and constructive problem discussion (κ = .62).

RMICS validity has been demonstrated in the following studies: Several RMICS codes are associated with dropping out of wife abuse treatment and for continued use of physical aggression posttreatment (i.e., predictive validity; Heyman, Brown, Feldbau, & O'Leary, 1999); most RMICS codes discriminate between distressed and nondistressed couples (i.e., discriminative validity; Vivian & Heyman, 1994[4]); RMICS code combinations were sensitive to experimental manipulations in an interpersonal closeness study in a theoretically predicted way (i.e., discriminative validity; Aron, Norman, Aron, McKenna, & Heyman, 2000).

Dyadic Adjustment Scale (DAS; Spanier, 1976)—The DAS, a 32-item self-report inventory designed to measure the severity of relationship discord in intimate dyads, was used in Sample 1. Scores range from 0 to 151, with higher values indicating more favorable adjustment; a score below 98 is traditionally considered to indicate clinical marital distress (e.g., Jacobson & Truax, 1991). The DAS has high convergent validity with other measures of marital adjustment and satisfaction (e.g., Heyman, Sayers, & Bellack, 1994). It had high internal consistency in Sample 1 (Cronbach's alpha = .95). Over 1,000 published studies have used the DAS.

Procedure

Observational data from both samples were collected using the standard marital problem-solving paradigm (e.g., Gottman, 1979). In Sample 1, couples chose a topic from the DAS that they had listed as generating frequent disagreements. The 15-minute marital interaction was part of a larger 3-hour assessment. All couples received $80 for participating; couples in the clinic subsample also had their intake fee ($35 to $70) waived. In Sample 2, topics for disagreement were selected by the researchers based on couples' responses to a list of 15 areas of disagreement (similar to items 1 to 15 of the DAS). The 15-minute interaction was part of a larger assessment protocol for which couples received $100 for completing two separate sessions of assessments, each session lasting about 1.5 hours.

Waters (1978) presented an approach for using standard psychometric reliability statistics to estimate the length of observation necessary to achieve adequate stability of codes. In the

[4]This conference presentation was based on preliminary results. The complete results are now displayed in Tables 1 and 2. Based on Tables 1 and 2, the following codes evidenced discriminative validity: husbands' psychological abuse, $t(188.00) = 3.67$, $p < 0.001$, distress-maintaining attributions, $t(216.61) = 5.50$, $p < 0.001$; hostility, $t(191.67) = 12.78$, $p < 0.001$; dysphoric affect, $t(201.18) = 3.64$, $p < 0.001$; withdrawal, $t(228.29) = 3.57$, $p < 0.00$; self-disclosure, $t(238.00) = -1.86$, $p < 0.03$; humor, $t(60.61) = -4.40$, $p < 0.001$; constructive problem discussion, $t(105.95) = -9.89$, $p < 0.001$; wives' psychological abuse, $t(188) = 3.93$, $p < 0.001$; distress maintaining attributions, $t(170.43) = 5.77$, $p < 0.001$; hostility, $t(212.85) = 11.22$, $p < 0.001$; dysphoric affect $t(188) = 1.79$, $p < 0.04$, withdrawal, $t(197.52) = 4.86$, $p < 0.001$; humor, $t(76.86) = -3.96$, $p < 0.001$, constructive problem discussion, $t(122.71) = -9.89$, $p < 0.001$.

psychometric theory of test reliability (Cronbach, 1951), one of the conventional methods to estimate reliability of tests is to assess 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 (the target behavior occurs or does not occur). In other words, an event-based coding system like the RMICS is converted into an interval-based coding system, using 30-second intervals.

Time intervals were 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. Thus, in this study, if there were 30 items (intervals) in each couple's interaction over the 15 minutes, two observations were created by dividing the 30 items (thirty 30-second sampling intervals) into two (i.e., odd and even) item groups. The correlation of the odd and even groups for the particular RMICS code reflects the internal consistency reliability.

Because the number of speaker turns was associated with the number of codes (r = .90), internal consistency reliability was adjusted for the total number of turns held by each individual. Specifically, for each time sample, the frequency of each code was divided by the total number of speaker turns. In addition, Spearman-Brown estimates of the duration of time sampling necessary to achieve conventional psychometric standards of reliability (r = .90) were calculated.[5] The Spearman Brown formula in this case is $.90 = kr_{sh}/[1 + r_{sh}(k-1)]$, where .90 is the target internal consistency reliability level we are trying to reach, k is the number of interval samples needed, and r_{sh} is the split half correlation estimate of the internal consistency reliability (for the code under consideration). This formula reduces to $k = 9[1/(r_{sh}-1)]$.

Results and Discussion

Tables 1, 2, and 3 present the results for Study 1's distressed and nondistressed subsamples and Study 2's engaged sample, respectively. For both husbands and wives, each table presents the mean base rate of occurrence (percent of total codes that a particular code constituted) and its standard deviation, along with the percent of participants who had at least one speaker or listener turn scored with a particular code. Next is the split half internal consistency reliability, adjusted for the total number of turns held by each individual. Reliability in this context indicates the stability of the observed data between odd and even 30-second interval samples. The final column is the estimated time necessary to collect stable data (i.e., split half r = .90).

For maritally distressed couples presenting for treatment (see Table 1), 15 minutes of observation is enough for almost all RMICS codes. It is barely too little to witness stable levels of husbands' relationship-enhancing attributions (16.7 minutes required). However, witnessing stable levels of husbands' and wives' self-disclosure would require 51.5 and 21.5 minutes, respectively, far more than most laboratory protocols could bear. Finally, husbands' dysphoric (i.e., sad) affect occurred in only 2.6% of participants and was too infrequent to even estimate length of time necessary for stability.

For happily married community couples (see Table 2), psychological abuse, dysphoric affect, and withdrawal all occur too infrequently to have stable estimates of frequency made in brief

[5]The pattern of results is consistent regardless of the reference reliability level (i.e., higher levels of desired reliability would require proportionally more time from all codes, lower levels of reliability would require less time). The conclusions of this paper are not a function of the chosen reference reliability level. The chosen reference reliability level (.90) is a conventional goal for correlational designs. Lower reliabilities can be adequate for between-group designs, where loss of power due to lower reliabilities can be offset by larger sample sizes. Higher reliabilities may be necessary if the goal of the research is to determine the amount of time necessary for individual clinical assessment.

laboratory interactions. Additionally, husbands' use of humor would require 16.7 minutes of observation. Note, however, that most researchers would hypothesize that psychological abuse, dysphoric affect, and withdrawal all occur far more frequently in distressed versus happily married couples. The low frequency of these behaviors in a between-groups study, therefore, would not be problematic.

Because behavior differs not only by couple status and topic (e.g., Christensen & Heavey, 1990; Wieder & Weiss, 1980) but also by developmental status (see Bradbury, 1998), we also analyzed the behaviors of an engaged sample (see Table 3). In other words, we wanted to explore if engaged persons (due to being typically younger, happier, and together for less time than their married counterparts) use behaviors from the RMICS palette with a different frequency profile than do married couples, thus requiring different lengths of observation to observe their behavioral frequencies reliably. Although 15 minutes of observation is sufficient for all RMICS behaviors for women—except for acceptance (which requires 18 minutes) and psychological abuse—engaged men's behaviors are so constrained that 15 minutes is sufficient to witness stable levels of hostility, relationship-enhancing attribution, self-disclosure, and constructive problem discussion only. A ll other RMICS behaviors are so infrequent that extending the conversations would do little to increase stability.

To summarize, using three different samples (clinic, nondistressed community, and engaged), we found that 15 minutes was enough to witness enough behavior to make reliable (i.e., internally consistent) estimations of the frequencies of most RMICS codes (i.e., hostility, distress-maintaining attributions, self-disclosure, relationship-enhancing attributions, humor, constructive problem discussion). Withdrawal and acceptance can be reliably estimated in distressed and nondistressed couples, respectively, indicating that they would be appropriate for tests of between-group significance. Finally, 10 minutes is sufficient for many codes of interest.

Note that these findings about reliable estimations can provide guidance about how much time is necessary to observe couples discussing a single topic, using the RMICS. It does not imply that a 15-minute conversation provides all the data necessary to understand a couple's communication style. Several studies, recently reviewed by Heyman (in press), have found substantial variability of couples' conversations across topics and across time. The current findings imply that *each* conversation observed should be about 15 minutes in length. Indeed, such a brief amount of time necessary for reliable base rate estimates may encourage more investigators to conduct studies that collect multiple observations, thus allowing further investigation of Couple X Situation interactions.

Because the reliability of observation data is a function of both (a) the length of observation and (b) the rate that particular behaviors occur, it seems prudent that the reliability (not just the interrater agreement) of individual behavior categories be routinely reported (cf. Mitchell, 1979). As noted earlier, low levels of reliability attenuate validity analyses, making it difficult to assess if, for example, the failure to find significant differences was due to insufficient observational lengths necessary to reliably estimate frequencies or to lack of substantive association between the variables. Furthermore, this paper was intended to encourage observational investigators to calculate estimates of minimal observational length. Although the findings regarding the adequacy of a 10- to 15-minute marital interaction applies to the RMICS, that length of time may not apply to other coding systems, especially if they comprise codes with low base rates (e.g., microanalytic systems such as the MICS).

Calculation of reliability coefficients (and estimates of the adequacy of the observational length used) is easily accomplished for studies that already employ interval sampling. Even sequences can be incorporated easily into interval sampling protocols by indicating the order that events

occurred within each interval. Event sampling is more commonly used in the marital and family observation fields. The same reliability information can be obtained from coding systems that use event sampling by including a time dimension (e.g., indicating the passage of every 30 seconds) on the coding sheet.

Finally, the ease in which "how much time is necessary" calculations can be made should entice behavioral investigators from a variety of content areas to publish such figures. By empirically investigating a factor that in most fields becomes reified through convention, investigators can conduct observational research that is both maximally efficient and maximally scientifically defensible.

References

Aron A, Norman C, Aron E, McKenna C, Heyman RE. Couples shared anticipation in novel and arousing activities and experienced relationship quality. Journal of Personality and Social Psychology 2000;78:273–284. [PubMed: 10707334]

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

Beck JG. Turning 30: Grown-up at last. Behavior Therapy 1999;30:1–4.

Block, J. (1977). Advancing the science of personality: Paradigm shift or improving the quality of research. In D. Magnusson & N. Endler (Eds.), *Personality at the crossroads: Current issues in interactional psychology* (pp. 37–64). Hillsdale, NJ: Lawrence Erlbaum.

Boyle DJ, Vivian D. Generalized versus spouse-specific anger/hostility and men's violence against intimates. Violence and Victims 1996;11:293–317. [PubMed: 9210274]

Bradbury, T. N. (Ed.). (1998). *The developmental course of marital dysfunction.* New York: Cambridge University Press.

Cascardi M, Vivian D. Context for specific episodes of marital violence: Gender and severity of violence differences. Journal of Family Violence 1995;10:265–293.

Christensen A, Heavey CL. Gender and social structure in the demand/withdraw pattern of marital conflict. Journal of Personality and Social Psychology 1990;59:73–81. [PubMed: 2213491]

Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika 1951;16:297–334.

Ehrensaft M, Vivian D. Spouses' reasons for not reporting existing marital aggression as a marital problem. Journal of Family Psychology 1996;10:443–453.

Epstein S. Aggregation and beyond: Some basic issues on the prediction of behavior. Journal of Personality 1983;51:360–392.

Foster DA, Caplan RD, Howe GW. Representativeness of observed couple interaction: Couples can tell, and it does make a difference. Psychological Assessment 1997;9:285–294.

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

Gottman, J. M. (1979). *Marital interaction: Experimental investigations.* New York: Academic Press.

Gottman JM. Consistency of nonverbal affect and affect reciprocity in marital interaction. Journal of Consulting and Clinical Psychology 1980;48:711–717.

Hahlweg, K., Reisner, L., Kohli, G., Vollmer, M., Schindler, L., & Revenstorf, D. (1984). Development and validity of a new system to analyze interpersonal communication: Kategorien-system für partnerschaftliche Interaktion KPI. In K. Hahlweg & N. S. Jacobson (Eds.), *Marital interaction: Analysis and modification* (pp. 182–198). New York: Guilford Press.

Hartmann, D. P., & Wood, D. D. (1990). Observational methods. In A. S. Bellack, M. Hersen, & A. E. Kazdin (Eds.), *International handbook of behavior modification and therapy* (2nd ed., pp. 107–138). New York: Plenum Press.

Heyman, R. E. (in press). Observation of couple conflicts: Clinical assessment applications, stubborn truths, and shaky foundations. *Psychological Assessment.*

Heyman RE, Brown PD, Feldbau SR, O'Leary KD. Couples' Communication Variables as Predictors of Dropout and Treatment Response in Wife Abuse Treatment Programs. Behavior Therapy 1999;30:165–190.

Heyman RE, Eddy JM, Weiss RL, Vivian D. Factor analysis of the Marital Interaction Coding System. Journal of Family Psychology 1995;9:209–215.

Heyman RE, Sayers SL, Bellack AS. Global Marital Satisfaction vs. Marital Adjustment: Construct validity and psychometric properties of three measures. Journal of Family Psychology 1994;8:432–446.

Heyman, R. E., & Vivian, D. (1993). *RMICS: Rapid Marital Interaction Coding System: Training Manual for coders.* Unpublished manuscript, State University of New York, Stony Brook, NY. (Available at http://www.psy.sunysb.edu/marital)

Heyman, R. E., Vivian, D., Weiss, R. L., Hubbard, K., & Ayerle, C. (1993, November). *Coding marital interaction at three levels of abstraction.* Paper presented at the 27th Annual Meeting of the Association for Advancement of Behavior Therapy, Atlanta, GA.

Heyman RE, Weiss RL, Eddy JM. Marital Interaction Coding System: Revision and empirical evaluation. Behavioural Research and Therapy 1995;33:737–746.

Hollingshead, A. B. (1975). *Four factor index of social status.* Unpublished manuscript, Yale University, Department of Sociology, New Haven, CT.

Holtzworth-Munroe A, Jacobson NS. Toward a methodology for coding spontaneous causal attributions: Preliminary results with married couples. Journal of Social and Clinical Psychology 1988;7:101–112.

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

Jacobson NS, Truax P. Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. Journal of Consulting and Clinical Psychology 1991;59:12–19. [PubMed: 2002127]

Kelly MB. A review of the observational data-collection and reliability procedures reported. Journal of Applied Behavior Analysis 1977;10:97–101. [PubMed: 16795549]

Kuhn, T. S. (1970). *The structure of scientific revolutions* (2nd ed.). Chicago: University of Chicago Press.

Margolin G, John RS, Gleberman L. Affective responses to conflictual discussion in violent and nonviolent couples. Journal of Consulting and Clinical Psychology 1988;56:24–33. [PubMed: 3346445]

Mischel, W. (1968). *Personality and assessment.* New York: Wiley.

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

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

Spanier GB. Measuring dyadic adjustment: New scales for assessing the quality of marriage and similar dyads. Journal of Marriage and the Family 1976;38:15–28.

Straus MA. Measuring intrafamily conflict and violence: The Conflict Tactics (CT) Scale. Journal of Marriage and the Family 1979;41:75–78.

Suen HK. Agreement, reliability, accuracy, and validity: Toward a clarification. Behavioral Assessment 1988;10:343–366.

Vivian, D., & Heyman, R. (1994, November). Aggression against wives: Mutual verbal combat "in context." In V. M. Follette (Chair), *Gender issues in couples research.* Symposium conducted at the 28th Annual Convention of the Association for Advancement of Behavior Therapy, San Diego, CA.

Vivian D, Langhinrichsen-Rohling J. Are bi-directionally violent couples mutually victimized? A gender-sensitive comparison. Violence and Victims 1994;9:107–123. [PubMed: 7696192]

Waters EB. The reliability and stability of individual differences in infant-mother attachment. Child Development 1978;49:483–494.

Waters, E., & Crowell, J. (1990). *Adult attachment models: Development after marriage.* Unpublished manuscript, Department of Psychology, State University of New York, Stony Brook, NY.

Weiss, R. L., & Heyman, R. E. (1997). Couple interaction. In W. K. Halford & H. J. Markman (Eds.), *Clinical handbook of marriage and couples intervention* (pp. 13–41). New York: Wiley.

Wieder GB, Weiss RL. Generalizability theory and the coding of marital interactions. Journal of Consulting and Clinical Psychology 1980;48:469–477. [PubMed: 7400432]

Appendix

How to Determine How Much Observational Data Are Necessary for Reliable Observations

1. (For event sampled data): Select a time interval (e.g., 15-sec., 30-sec.) that is longer than the typical duration of the behaviors of interest. Shorter intervals may be appropriate if the behavior(s) of interest occur in bursts, rather than consistently across the observation.
2. (For event sampled data): Indicate the passage of these intervals throughout the collection of the event sampled observation.
3. When data collection is complete, divide each subject's observations into separate sets of odd and even time intervals. Determine the number of times the target code occurred in odd intervals and even intervals.
4. Create a variable for the frequency of occurrence of the target variable during the odd intervals (e.g., for code A, name variable oddA).
5. Create a variable for the frequency of occurrence of the target variable during the even intervals (e.g., for code A, name variable evenA).
6. Correlate oddA with evenA to obtain r_{sh} (i.e., the reliability obtained in each of the halves of the full observation period). The reliability for the full observation is $rel = kr_{sh} / (1 + r_{sh}(k - 1))$, which for one set of odd and one set of even observations (i.e., $k = 2$), is $rel = (2\ r_{sh}) / (1 + r_{sh})$.
7. Calculate the extent to which the amount of time observed to would have to increase (or could decrease, if desired) to obtain a specified level of reliability (e.g., .90):

$$\text{Observation Duration Multiplier} = \frac{\text{desired reliability } (1 - \text{obtained reliability})}{\text{obtained reliability } (1 - \text{desired reliability})}$$

   For example, given 15 minutes of observation, an obtained reliability of .6, and a desired reliability of .9:

$$\text{Observation Duration Multiplier} = \frac{.9(1 - .6)}{.6(1 - .9)} = \frac{.36}{.06} = 6$$

   This indicates that one would have to observe for 6 × 15 minutes (or 90 minutes) to estimate individuals' typical rate of code A with a reliability of .9.
8. Repeat steps 3 to 7 for each code of interest.
9. Information of this type from pilot data can be used to modify a proposed protocol. Investigators should ensure that the length of observation is sufficient even for codes that are very unevenly distributed across time or that have low base rates (i.e., codes with the highest multipliers).
10. Note that similar calculations can be made to determine if lengthy observations can be reduced. For example, given 30 minutes of observation, an obtained reliability of .9, a researcher whose design could accept a reliability of .8 would find:

$$\text{Observation Duration Multiplier} = \frac{.8(1-.9)}{.9(1-.8)} = \frac{.08}{.18} = .44$$

This indicates that the observation time could be reduced to 30 minutes × .44 (or 13.2 minutes) with only a modest reduction in reliability.

TABLE 1

Distressed Subsample: RMICS Descriptive Data, Interval-Sampled Reliabilities, and Estimated Times Needed to Achieve Adequate Reliabilities

	Husbands					Wives				
Code	***M*** **(%)**	***SD***	**% Participants**	**Reliability**	**Estimated Time (min.)**	***M*** **(%)**	***SD***	**% Participants**	**Reliability**	**Estimated Time (min.)**
PA	0.6	2.1	12.2	.98***	3	0.4	1.5	10.2	.92***	11.7
DA	2.5	3.1	61.2	.92***	11.7	4.6	6	71.9	.94***	8.4
HO	34.4	27.5	89.3	.96***	6	40.9	27.2	92.9	.97***	4.5
DY	0.2	1.4	2.6	n/f	n/f	0.8	2.8	12.8	.99***	1.5
WI	1.1	2.8	27	.95***	6.6	0.7	2.2	16.3	.96***	6
RA	3.5	4.1	72.5	.89***	16.7	3.2	4.2	65.8	.93***	10.1
AC	1.1	2.8	26	.73***	51.5	0.5	1.4	18.4	.86***	21.6
SD	7.9	9.8	81.6	.98***	3	6.9	7.7	80.6	.93***	10.1
HM	2.3	5.1	42.9	.91***	13.4	2.1	4.2	42.9	.91***	13.4
PD	47.3	24	99.5	.95***	6.6	40.9	23.9	98.5	.95***	6.6

Note. $n = 196$

*** $p < .001$; n/f: behavior was too infrequent to estimate for reliability. PA = psychological abuse; DA = distress-maintaining attributions; HO = Hostility; DY = Dysphoric Affect; WI = Withdrawal; RA = Relationship-Enhancing Attributions; AC = Acceptance; SD = Self-Disclosure; HM = Humor; PD = Constructive Problem Discussion.

TABLE 2

Nondistressed Subsample: RMICS Descriptive Data, Interval-Sampled Reliabilities, and Estimated Times Needed to Achieve Adequate Reliabilities

	Husbands					Wives				
Code	***M***	***SD***	**% Participants**	**Reliability**	**Estimated Time (min.)**	***M***	***SD***	**% Participants**	**Reliability**	**Estimated Time (min.)**
PA	0	0	0	n/o	n/o	0	0	0	n/o	n/o
DA	0.7	1.4	22.5	.82***	30	1.7	2.2	51	.94***	8.4
HO	6.7	10.9	54.9	.97***	4.5	9.3	13.6	66.7	.95***	6.6
DY	0	0	0	n/o	n/o	0	0.3	2	n/f	n/f
WI	0.1	0.4	7.8	.69**	60	0.1	0.4	3.9	47**	144
RA	3.1	4.1	54.9	.95***	6.6	3.7	4.3	64.7	.93***	10.1
AC	2.0	5.3	41.2	.95***	6.6	1.1	2.7	29.4	.88***	18
SD	7.8	7.9	88.2	.93***	10.1	9.0	8.7	86.3	.94***	8.4
HM	5.3	5.2	72.5	.89***	16.7	6.2	6.6	74.5	.92***	11.7
PD	74.5	15.5	98	.92***	11.7	69.4	18.2	98	.95***	6.6

Note. $n = 51$

** $p < .01$

*** $p < .001$; n/f: Behavior was not frequent enough to estimate reliability; n/o: The particular behavior was not observed. PA = Psychological Abuse; DA = Distress-maintaining attributions; HO = Hostility; DY = Dysphoric Affect; WI = Withdrawal; RA = Relationship-Enhancing Attributions; AC = Acceptance; SD = Self-Disclosure; HM = Humor; PD = Constructive Problem Discussion.

TABLE 3

Engaged Sample: RMICS Descriptive Data, Interval-Sampled Reliabilities, and Estimated Times Needed to Achieve Adequate Reliabilities

	Husbands					Wives				
Code	***M***	***SD***	**% Participants**	**Reliability**	**Estimated Time (min.)**	***M***	***SD***	**% Participants**	**Reliability**	**Estimated Time (min.)**
PA	0.1	0.7	3.5	n/f	n/f	0.2	1.9	2.1	n/f	n/f
DA	1	1.6	43	.79***	35.6	1.3	2	49.3	.90***	15
HO	13	16	78.9	.94***	8.4	14.2	16.8	80.3	.92***	11.7
DY	0.1	1	2.1	n/f	n/f	0.4	2.5	5.6	.98***	3
WI	0.3	2.5	4.2	n/f	n/f	0.1	0.3	2.8	.67**	65
RA	2.8	3.3	76.8	.91***	13.4	2.9	2.8	74.6	.94***	8.4
AC	0.3	0.8	16.2	.43**	191.3	0.3	0.9	11.3	.88***	18
SD	2.9	4.8	60.6	.95***	6.6	2.9	4.3	66.2	.97***	4.5
HM	2.9	3.8	61.3	.82***	30	2.9	4.2	58.5	.88***	18
PD	71.2	20.2	100	.96***	6	69.5	19.9	100	.97***	4.5

Note. $n = 142$

** $p < .01$

*** $p < .001$; n/f: Behavior was not frequent enough to estimate reliability. PA = Psychological Abuse; DA = Distress-maintaining attributions; HO = Hostility; DY = Dysphoric Affect; WI = Withdrawal; RA = Relationship-Enhancing Attributions; AC = Acceptance; SD = Self-Disclosure; HM = Humor; PD = Constructive Problem Discussion.