Published in final edited form as:
Psychol Assess. 2012 December ; 24(4): . doi:10.1037/a0028680.

Do the Naïve Know Best? The Predictive Power of Naïve Ratings of Couple Interactions

Katherine J.W. Baucom,
Department of Psychology, University of California, Los Angeles

Brian R. Baucom, and
Department of Psychology, University of California, Los Angeles

Andrew Christensen
Department of Psychology, University of California, Los Angeles

Abstract

We examined the utility of naïve ratings of communication patterns and relationship quality in a large sample of distressed couples. Untrained raters assessed 10-minute videotaped interactions from 134 distressed couples who participated in both problem solving and social support discussions at each of three time points (pre-therapy, post-therapy, and 2-year follow-up) during a randomized clinical trial of behavioral couple therapy. Teams of naïve raters observed a particular type of discussion from the three time points at one sitting in a random order and rated dyadic interaction patterns (negative reciprocity, positive reciprocity, wife demand/husband withdraw, husband demand/wife withdraw, and mutual avoidance) and the overall relationship quality of couples. These naïve ratings were strongly and consistently associated with both levels of, and changes in, trained observational codes and self-reported relationship satisfaction. Naïve ratings of couples accounted for similar – and at times superior – amounts of variance in both concurrent relationship satisfaction and divorce at 5-year follow-up when compared with trained ratings. These findings offer compelling support for the use of naïve raters in research with couples, and also suggest important future directions that are applicable to both research and practice with distressed couples.

Keywords

behavioral assessment; couples; observation methods; interpersonal communication; relationship satisfaction

Virtually all theoretical perspectives consider communication to be a critical element in relationship functioning; there is a wealth of empirical evidence that supports links between the ways couples communicate and the quality of their relationship (Weiss & Heyman, 1997). The most common approach for objective assessment of couples' communication is observational coding by highly trained coders (see Heyman, 2001 for a review). Although

Correspondence concerning this article should be addressed to Katherine J.W. Baucom, M.A., 1285 Franz Hall, Box 951563, Los Angeles, CA 90095-1563; 310-430-4349 (voice), 310-206-5895 (fax); kwilliams@ucla.edu.
Brian R. Baucom is now in the Department of Psychology at University of Southern California.

important information has been gained from this method, the approach has a number of limitations and restrictions, including the need for substantial resources, high reliability at times trumping "ecological validity," and difficulty with replication across sites and coding systems. The current study builds upon recent developments in communication assessment suggesting that naïve[1] or minimally trained raters can reliably and validly code interactions in couples and families at a high level of agreement with traditional observational methods (e.g., Baker, Haltigan, Brewster, Jaccard, & Messinger, 2010; Lorber, 2006; Waldinger, Schulz, Hauser, Allen, & Crowell, 2004). This nascent body of research suggests that further development and additional application of these techniques may yield findings unattainable with traditional methods.

Limitations to Traditional Observational Coding Systems

Both macro- and microanalytic observational coding systems are used in research with couples. In the former, coders are asked to make a judgment of the extent to which partners exhibit behaviors of interest over the course of an entire discussion, whereas in the latter they typically code either the frequency or intensity of target behaviors in segments of the discussion usually ranging from a talk turn to a 30-second interval. Despite strong contributions to the study of relationships, both types of observational systems have important weaknesses. In both micro- and macro-analytic systems, comparing findings from observational studies can be problematic. The way one research group operationalizes negative behavior may be quite different from another yet findings using different definitions of a construct are compared across research groups. Even at the same site with the same system, there can be coder drift from one team of coders to another. Research findings are also compared across types of systems when the information gleaned from them is actually quite different. For example, coding the frequency with which a person uses negative behaviors in a microanalytic system is not necessarily consistent with the extent to which a partner is negative over the course of a discussion in a macroanalytic system.

Microanalytic systems such as the Marital Interaction Coding System-IV (MICS-IV; Heyman, Weiss, & Eddy, 1995) and the Specific Affect Coding System (SPAFF; Gottman, McCoy, Coan, & Collier, 1996) are particularly burdensome for researchers given the time coding and training takes. Some traditional microanalytic systems required coders to pause and code every change in affect from neutral, although more recent developments of these systems code partners every second (e.g., SPAFF; Coan & Gottman, 2007) or talk turn (MICS-IV). Although they allow for detailed behavior analysis, such systems are prone to low coder reliability, especially with more complex codes. Coding short intervals also does not allow coders to take into account important contextual information in adjacent ratings (e.g., the sarcastic tone that was used after a partner smiled sweetly). Furthermore, the low frequency of codes creates the need to collapse either into larger codes or across couples, which takes away important distinctions in behavior. Last, microanalytic systems can take an exorbitant amount of time, ranging from 1.5 hours (Heyman, 2004) to 20 hours (Gottman, 1994) to code a 10–15 minute discussion.

A larger concern of observational systems in general is their ecological validity. Researchers have argued that microanalytic coding systems in particular take away the ability for coders to make intuitive judgments since they are restricted by the rules of observational systems (e.g., Lorber, 2006; Waldinger et al., 2004). Although macroanalytic systems offer coders more room to judge partners' communication over the course of a discussion and are less time-consuming, they are not without limitations. A growing body of research suggests that

[1]We use the term "naïve raters" throughout this paper to describe individuals who are naïve to research with couples in general and observational rating systems in particular.

adhering to a researcher's conceptualization of constructs of interest may hinder raters' ability to – as Supreme Court Justice Potter Stewart said of pornography – "know [certain communication behaviors] when [coders] see it" even if that specific behavior has not been included in the researcher's definition. Although this focus certainly helps maintain reliability among coders, it may be at the expense of validity since it restricts raters' use of their own judgment. A related limitation is that macroanalytic coding systems generally don't allow for examination of sequences of behavior that are strongly linked with relationship outcomes and that clinicians can easily observe in working with couples.

Intuitive Judgments of Relationship Interactions

Recent empirical studies have given raters flexibility in their assessment of communication in couples (i.e., in an "ecologically valid" manner). Results from such studies suggest that naïve raters, who are not given detailed descriptions of constructs or trained in how to rate them but are asked to rely on their intuitive understanding of those constructs, are able to reliably rate communication and emotional expression in couples and that these ratings correlate with those of trained raters. Waldinger and colleagues found that untrained observers' ratings of items based on the SPAFF loaded on four distinct factors: hostility, distress, affection, and empathy (Waldinger et al., 2004). These factors were correlated with concurrent self-reported and interviewer-rated relationship quality and, to some extent, likelihood of divorce after 5 years. To establish validity, untrained observers in this study also rated discussions of four couples that had been previously coded with the SPAFF in a separate sample. Untrained ratings were correlated with SPAFF coding of the same discussions at the level of the conversation (i.e., aggregated over 30-second segments), but correlations between the two systems were low to moderate at 30-second segments. The comparison was somewhat difficult, however, since the SPAFF codes the frequency of affect although the untrained system assessed the intensity of affect. Although the examination of untrained ratings in the larger sample of couples is encouraging, the sample size was still relatively small ($N = 47$) and included ratings of individual affect but not dyadic interaction patterns. Additionally, the comparison between the naïve system and the SPAFF was in too small a number of couples to examine whether intuitive ratings actually add to the predictive power of trained codes, an important next step in this area of research.

Adding to the research on intuitive ratings of communication, Ebling and Levenson (2003) asked those with personal experience (e.g., happily and unhappily married individuals), those with professional experience (e.g., marital researchers, clinical psychology graduate students), and undergraduates to whom marriage was not personally or professionally relevant to make ratings of 3-min segments of videotaped discussions between partners. There were no significant differences in the ability of these ratings to predict relationship stability but those with personal experience were significantly more accurate than those with professional experience in predicting relationship satisfaction; undergraduates were not included in these comparisons. Several other studies have found empirical support for the validity of untrained ratings of mothers' overreactive discipline (Lorber, 2006) and both infant and parent emotion (Baker et al., 2010). Taken together, the findings of these studies lend support to the notion that lay people have intuitive knowledge about behavior and affect that can be accessed without formal training, although it is unclear whether the information naïve ratings provide offers additional unique information beyond that captured with trained ratings.

Key Dyadic Interaction Patterns for Relationship Functioning

Perhaps the signature of a distressed couple is negative reciprocity, where one partner expresses negativity and the other responds in kind. Distressed couples get caught in

negative cycles more often and for longer periods of time than do nondistressed couples (Snyder, Heyman, & Haynes, 2005), leading researchers to conclude that "negativity is like a black hole for distressed couples (Weiss & Heyman, 1997, p. 23)." Another widely researched dysfunctional interaction pattern is demand/withdraw, where one partner demands or pressures for change and the other withdraws from the discussion. This pattern is strongly and consistently associated with relationship dissatisfaction across levels of distress (Eldridge, Jones, Sevier, Atkins, & Christensen, 2007), sexual orientation (Baucom, McFarland, & Christensen, 2010), and culture (Christensen, Eldridge, Catta-Preta, Lim, & Santagata, 2006). Similarly, mutual avoidance is exhibited more frequently in distressed couples than their nondistressed counterparts (e.g., Noller & White, 1990) although there are some findings to the contrary (Gottman, 1994). Although substantially less research has been devoted to healthy or functional dyadic interaction patterns, there is some evidence that more positive reciprocity is associated with greater relationship satisfaction in couples (Margolin & Wampold, 1981).

As outlined, researchers have primarily used observational systems to examine individual partners' behavior typically under the umbrella of negativity (e.g., contempt, anger) and positivity (e.g., affection, compromise) (Weiss & Heyman, 2004). Although microanalytic systems allow for examination of dyadic sequences such as positive reciprocity, there are a number of limitations with using this approach. The use of naïve ratings may be an ideal way to examine patterns of interaction in couples, particularly patterns that are less straightforward and would be difficult to recognize using microanalytic coding methods. For example, if partner A criticizes, partner B initially responds neutrally but then changes the direction of the conversation it would not likely be seen as a sequence from a microanalytic perspective but a naïve rater might easily identify this sequence as demand/withdraw.

In summary, traditional observational methods of assessment with couples have a number of limitations, many of which can be addressed through more flexible and intuitive methods of naïve observational rating. With growing support for the use of untrained or "naïve" ratings of couples' interactions, more research is needed to determine whether such ratings not only provide similar information to trained ratings but actually provide *additional* information compared to trained ratings. In this study naïve raters specifically assessed several dyadic interaction patterns with well-documented links with relationship functioning: negative reciprocity, demand/withdraw, mutual avoidance, and positive reciprocity.

Current Study Aims

We examine two aspects of naïve ratings of distressed couples from a large randomized clinical trial of behavioral couple therapies during two different types of discussions (problem solving and social support) at three time points (pre-therapy, post-therapy, 2-year follow-up). First, we examine associations between naïve ratings and trained codes for similar behaviors exhibited during the same discussions. Second, we examine associations between naïve ratings and relationship outcomes by testing naïve ratings' overall predictive power as well as their predictive power in comparison to trained ratings.

Hypothesis 1

Naïve observational ratings of communication will be associated with (1a) concurrent trained observational codes and (1b) concurrent relationship satisfaction. We expect that greater naïve ratings of relationship quality and positive reciprocity will be associated with greater relationship satisfaction and higher scores of trained positive communication codes (i.e., positivity and problem solving) and lower scores of trained negative communication codes (i.e., negativity, withdrawal); we expect the reverse will be true of the dysfunctional

interaction patterns (i.e., negativity reciprocity, wife demand/husband withdraw [WD/HW], husband demand/wife withdraw [HD/WW], mutual avoidance).

Hypothesis 2

Changes over time in naïve ratings of communication will be associated with changes in both (2a) trained codes of communication and (2b) relationship satisfaction. More specifically, we expect that increases in naïve ratings of relationship quality and positive reciprocity, and decreases in the naive ratings of dysfunctional interaction patterns, will be linked with increases in relationship satisfaction and trained positive codes and decreases in trained negative codes.

Hypothesis 3

Naïve ratings will provide unique information about relationship functioning compared with trained codes of similar constructs. In the prediction of concurrent relationship satisfaction and long-term relationship stability, we compare naïve ratings of relationship quality to all trained codes; naïve ratings of negative reciprocity to trained codes of negativity; naïve ratings of positive reciprocity to trained codes of positivity; naïve ratings of mutual avoidance to trained codes of withdrawal; and naïve ratings of HD/WW and WD/HW to trained codes of these constructs.

Method

Participants

Participants were 134 seriously and chronically distressed married couples that participated in a randomized clinical trial of two behaviorally based couple therapies conducted at [Site 1] and [Site 2]. Couples were recruited for the [Site1/Site2] Couple Therapy Project through media ads and flyers promoting free therapy.

Couples included in the larger study scored in the distressed range on three measures of relationship satisfaction administered at three different time points, and were excluded from the study if they were not married, did not speak English, or reported moderate to severe domestic violence. Based on the combination of pre-treatment relationship satisfaction from two self-report measures of relationship satisfaction, couples were classified as either moderately or severely distressed. Within each level of distress, couples were randomly assigned to participate in either TBCT or IBCT (see [Author] for a more detailed description of recruitment, screening, and stratification procedures). The mean ages of husbands and wives were 43.5 years (*SD* = 8.8) and 41.6 years (*SD* = 8.6) respectively. The mean years of marriage were 10.0 (*SD* = 7.6). The mean number of children that couples had was 1.1 (*SD* = 1.0). The sample was 77.6 % Caucasian, 7.5 % African American, 5.2 % Asian American/Pacific Islander, 5.2 % Latino, .6 % Native American/Alaskan Native, and 4.1 % other.

Measures

Relationship Satisfaction—The Dyadic Adjustment Scale (DAS; Spanier, 1976) was used to measure relationship satisfaction at each time point in the current investigation. Within this sample, subscale reliabilities were good for both husbands' and wives' scores (αs = .89 and .87, respectively). The sample represented couples that were seriously and chronically distressed (husbands' DAS *M* = 84.49, *SD* = 14.96; wives' DAS *M* = 84.70, *SD* = 13.98).

Trained Observational Coding—Behavior during the 10-minute discussions was coded using the Couple Interaction Rating System (CIRS; [Author]) and the Social Support

Interaction Rating System (SSIRS; [Author]). The CIRS is a 13-item observational coding system ([Author]) designed to capture problem solving and communication behaviors during problem solving discussions, and the SSIRS is an 18-item observational rating system designed to capture emotional features of the interaction ([Author]). Both systems are global coding systems designed to capture overall impressions of behaviors in context, rather than specific counts of behaviors or specific use of any particular language. Both included items coded on a Likert scale of 1 (*none*) to 9 (*a lot*).

Coders were instructed to focus on one spouse at a time, and make judgments about the extent to which the target spouse demonstrated the behavior specified by the codes (using information from both verbal and nonverbal behaviors) during the discussion. Coders considered the frequency, context, and intensity of total interaction behavior of each partner in each discussion. Coders were blind to all hypotheses, and coded pre-therapy, post-therapy, and 2-year follow up discussions in a random order. Multiple coders on multiple coding teams rated the discussions, but a number of steps were taken to ensure consistency across teams. Coders received didactic training in the coding system and practiced on training tapes until they were reliable. Once trained, coders participated in more than one team, coded both distressed and nondistressed couples, and met weekly for reliability meetings where they received additional instructions from the research supervisors on how to code items. See [Author] for a more detailed description of coding procedures and controls.

In the previous paper on pre- and post-therapy data from the larger study, principal component analyses revealed four components: negativity, withdrawal, positivity, and problem solving ([Author]). Subscale reliabilities among scores across all time points included in the present paper were: negativity (α = .91), withdrawal (α = .76), positivity (α = .72), and problem solving (α = .72). Interobserver reliabilities of scores were generally good for negativity (αs = .86 to .95), withdrawal (αs = .79 to .88), positivity (αs = .81 to .95), and problem solving (αs = .66 to .92). Trained ratings of demand/withdraw patterns in each direction were computed by summing one partner's demand behavior (i.e., the average of ratings of blame and pressures for change) and the other's withdrawal from a given interaction, consistent with previous studies using these coding systems (Baucom et al., 2010; Eldridge et al., 2007).

Naïve Observational Rating—The Naïve Observational Rating System (NORS; [Author]) is a 15-item global observational rating system that we developed to capture communication during couples' interactions. Raters were given a general description of each construct they rated.[2] We focus only on the 6 ratings included in this paper. Relationship quality was coded on a 100-point scale (higher scores representing greater quality of the relationship). The following five dyadic interaction patterns were rated on a Likert scale of 1 (*low*) to 10 (*high*): negative reciprocity, positive reciprocity, WD/HW, HD/WW, and mutual avoidance. In the NORS, raters were instructed to focus on both partners at a time, and make judgments about the extent to which the couple demonstrated the dyadic patterns specified by the ratings. Consistent with the CIRS and SSIRS, raters considered the frequency, context, and intensity of both partners' interaction behavior in their overall ratings of each discussion. Raters were blind to all hypotheses, the larger research study, and the area of close relationships in general; they were recruited for this project in particular and were not selected if they had research or course experience in the area of relationships. Raters met weekly to discuss ratings on which they disagreed, but were encouraged to come to a

[2]As an example, raters were given the following description of negative reciprocity: "To what extent did the couple exchange negative comments and negative nonverbal behavior in a "tit-for-tat" like way (e.g., criticize each other or exchange sarcastic comments, put-downs, frowns, sneers, or looking away in anger or disgust)?"

common understanding as a group as opposed to being instructed by the research supervisor on how to code constructs of interest.

Six paid undergraduate research assistants rated problem solving discussions, and five volunteer undergraduate research assistants subsequently rated the social support discussions. The raters of problem solving discussions read over the entire transcript of the discussion to get a sense of it, watched the discussion, and then rated it on the aforementioned dimensions. The raters of social support discussions did not read the transcript prior to coding, but rather just watched the discussion and then rated it. Within the two types of discussions, all discussions on topics chosen by the wife in our sample of couples were rated prior to discussions on topics chosen by the husband. Each couple's discussions of a given type were rated in one sitting (e.g., wife's topic problem solving discussions from pre, post, and 2-yr follow-up for a given couple), but the order of time points was randomized. Ratings of NORS items on the three discussions were judged independently with the exception of the relationship quality rating; raters had to give *different* scores on the quality of the relationship based on each of the three discussions they watched to allow for examination of changes in, and the order of, relationship quality in discussions rated (analyses beyond the scope of this manuscript). See Table 1 for descriptive statistics, reliabilities, and intercorrelations. Although a number of these ratings were highly correlated, we analyzed them separately since they are conceptually distinct.

Procedure

At each of three time points (pre-therapy, post-therapy, and 2-year follow-up), these 134 couples completed questionnaires in the lab. Post-therapy assessments occurred approximately 26 weeks after the pre-therapy assessment, when couples were at or near the end of treatment. They then participated in two 10-minute problem solving discussions, completed additional questionnaires, and then participated in two 10-minute social support discussions; all discussions were videotaped for future ratings. Couples discussed one relationship problem and one personal problem selected by the husband and one relationship problem and one personal problem selected by the wife, with the order of topics counterbalanced within each type of discussion.

Problem solving discussions included 133 couples at pre-therapy, 117 couples at post-therapy, and 84 couples at 2-year follow up. The social support discussions included 96 couples at pre-therapy, 87 couples at post-therapy, and 76 couples at 2-year follow up.[3]

Results

We used Hierarchical Linear Modeling 6.04 (HLM; Raudenbush, Bryk, Cheong, & Congdon, 2007) and Stata 11 (StataCorp, 2009) for analyses since both include multilevel models that account for the dependencies in data where time points are nested within couples.

Associations between Naïve Ratings and Concurrent Relationship Outcomes (Hypothesis 1)

Hypothesis 1 analyses—Our first hypothesis was that naïve ratings would be associated with both trained ratings (Hypothesis 1a) and relationship satisfaction (Hypothesis 1b) at each time point. To test this we ran a series of models in HLM predicting couple-level

[3]Although there were 134 couples in the larger study, there were a number of factors contributing to discussions not being available for all couples at all three time points. The UW site did not collect social support discussions initially, there were some problems with equipment or recordings, some couples did not fully participate in the procedure because of separation or divorce, although others dropped out of the study for various reasons.

relationship variables from naïve codes, controlling both for the difference between husbands and wives in the relationship variable being predicted and for time (*pre-therapy* = 0, *post-therapy* = 1, *2-year follow-up* = 2).[4] Within each model where relationship satisfaction was the outcome we included separate predictors for the naïve rating in problem solving and in social support discussions; this allowed us to examine unique effects of behavior in the context of these two types of discussions when predicting relationship satisfaction. We ran separate models with each of the four trained communication codes and relationship satisfaction being predicted by each of the six naïve ratings. Below is the exact model we ran for trained codes:

$$\begin{aligned}\text{Couple Average Trained Code}_{ij} = {} & \beta_0 + \beta_1(\text{Difference in Trained Code}) \\ & + \beta_2(\text{Naïve Rating}) \\ & + \beta_3(\text{Time}) \\ & + \beta_4(\text{Type}) \\ & + \beta_5(\text{Topic}) + r_{ij}\end{aligned} \quad (1)$$

We included a random effect on the intercept at Level 2 but did not include any Level-2 covariates. We use a Bonferroni correction to adjust significance levels for familywise error given the large number of models.

Hypothesis 1 results—As hypothesized, naïve ratings were strongly associated with trained codes of the same discussions (Hypothesis 1a; see Table 2). Higher levels of naïve ratings of both relationship quality and positive reciprocity were associated with higher trained scores on positive behaviors (i.e., positivity, problem solving; *p*s < .001) and lower trained scores on negative behaviors (i.e., negativity, withdrawal; *p*s < .001). Similarly, higher levels of the naïve ratings of the dysfunctional interaction patterns of negative reciprocity, HD/WW, and WD/HW were associated with lower trained scores of positive behaviors and higher trained scores of negative behaviors (*p*s < .001). Mutual avoidance was only significantly associated with greater withdrawal (*p* < .001) behavior.

We also found predicted associations between naïve ratings and concurrent relationship satisfaction (Hypothesis 1b; see Table 3). Higher naïve ratings of relationship quality and positive reciprocity in the problem solving discussions were uniquely associated with higher levels of self-reported relationship satisfaction (i.e., over and above effects of these ratings from social support discussions; *p*s < .001); higher naïve ratings of relationship quality in the social support discussions had a marginally significant association with higher levels of self-reported relationship satisfaction, and ratings of positive reciprocity in social support discussions was not significantly associated with self-reported relationship satisfaction. We found a similar pattern of results with the dysfunctional interaction patterns: higher levels of naive ratings of negative reciprocity, HD/WW, and WD/HW in the problem solving discussions were associated with lower self-reported relationship satisfaction (*p*s < .001). Greater HD/WW in the social support discussions was marginally associated with lower self-reported relationship satisfaction (*p* < .10) but neither negative reciprocity nor WD/HW in the social support discussions were significantly associated with lower self-reported relationship satisfaction as we had expected (*p*s > .10). Although greater mutual avoidance in problem solving discussions was not significantly associated with lower self-reported

[4]In predicting trained ratings we also controlled for interaction type (problem solving, social support) and whose topic was being discussed (husband's topic, wife's topic); we did not include type or topic in models predicting relationship satisfaction since satisfaction does not vary by type or topic at a given time point.

relationship satisfaction, we did find this effect with mutual avoidance in the social support discussions ($p < .001$).

Associations between Changes in Naïve Ratings and Changes in Relationship Outcomes (Hypothesis 2)

Hypothesis 2 analyses—Our second hypothesis was that changes in naïve ratings over time would be associated with changes in both trained codes (Hypothesis 2a) and relationship satisfaction (Hypothesis 2b). To test this we ran a series of cross-lagged regression models as outlined by Kenny, Kashy, and Cook (2006). Using Stata, we ran separate simultaneous models for changes in each of the six naïve ratings predicting changes in each of the trained codes and self-reported relationship satisfaction, and controlled for Time (i.e., time point being predicted). We centered all variables other than Time, which was dummy coded. The effects of interest were the slopes of the naïve ratings at Time(t) since they represent effects of the naïve ratings at that time point on the relationship outcome after controlling for both the naïve ratings and the relationship outcome from the previous time point, Time(t-1) (i.e., associations between changes in naïve rating and changes in relationship outcome). Equation 2 presents the full model.

$$\begin{aligned}\text{Average Couple Outcome(t)}_{ij} = {} & \beta_0 + \beta_1(\text{Average Couple Outcome}[\text{t}-1]) \\ & + \beta_2(\text{Naive Rating}[\text{t}-1]) + \beta_3(\text{Naïve Rating}[\text{t}]) \\ & + \beta_4(\text{Time}) + r_{ij}\end{aligned} \tag{2}$$

Hypothesis 2 results—We also found strong support for links between changes in naïve ratings of communication and changes in relationship variables, as presented in Table 4 (Hypothesis 2a). Increases in rated relationship quality and positive reciprocity were associated with increases in trained codes of positive behaviors and decreases in negative behaviors over time (*p*s < .001). The reverse was true for the dysfunctional interaction patterns of negative reciprocity, HD/WW, and WD/HW: increases in these rated communication patterns were associated with decreases in trained codes of positive behaviors and increases in negative behaviors. Increases in rated mutual avoidance were only significantly associated with increases in trained withdrawal ($p < .01$).

Table 5 displays a similar pattern of results (Hypothesis 2b); changes in naive ratings in both types of discussions were associated with changes in self-reported relationship satisfaction, although effects of ratings from problem solving discussions were more consistent than effects of ratings from social support discussions. As expected, increases in rated relationship quality and positive reciprocity were associated with increases in self-reported relationship satisfaction and there were unique effects of these ratings from each of the types of discussions (*p*s < .001). Increases in negative reciprocity, WD/HW, and HD/WW were associated with decreases in self-reported relationship satisfaction but only ratings from the problem solving discussions evidenced unique effects (*p*s < .05). Increases in mutual avoidance in the social support discussions – but not the problem solving discussions – were associated with decreases in self-reported relationship satisfaction ($p < .001$).

Prediction of Relationship Satisfaction and Relationship Status from Naïve Ratings versus Trained Codes of Communication (Hypothesis 3)

Hypothesis 3 analyses—Finally, we examined whether naïve ratings account for additional variance in self-reported relationship satisfaction and long-term relationship status when compared with trained codes of similar constructs. To test this we ran a series of

models in Stata examining: (1) whether naïve ratings alone accounted for more variance in concurrent relationship satisfaction and relationship stability at 5-year follow-up when compared with trained codes alone; (2) whether naïve ratings accounted for *additional* variance in relationship satisfaction and status when added to models with trained codes predicting these outcomes, and vice versa. To test whether separate (i.e., non-nested) models with naïve ratings and trained codes accounted for significantly different variance in relationship satisfaction we ran a series of Vuong (1989) tests for differences in non-nested models using a Stata program. Although there is not a significance test for differences in model fit between non-nested binomial logistic regression models, the models' Akaike's Information Criterion (AIC) values can be compared with differences of 2 or greater representing reliable differences in model fit (e.g., Burnham & Anderson, 1998, p.70). Finally, to examine unique effects (controlling for shared variance) of the naïve ratings and trained codes in predicting concurrent relationship satisfaction and subsequent divorce we ran likelihood ratio tests in Stata to determine whether the addition of one set of variables significantly increased the variance explained in the outcome variable by the other (i.e., a stepwise approach).

Hypothesis 3 results—Table 6 presents results from a series of non-nested (i.e., independent) models with naive ratings versus trained codes predicting concurrent relationship satisfaction, as well as divorce at 5-year follow-up. We found variance in relationship satisfaction to be similarly explained by naïve models relative to trained models with one exception: naïve ratings of negative reciprocity accounted for significantly more variance in relationship satisfaction than did trained codes of negativity ($p = .013$). Models that compared a) naïve ratings of positive reciprocity and positivity; b) naïve ratings of mutual avoidance and withdrawal; c) naïve ratings of WD/HW and the respective trained codes; d) naïve ratings of HD/WW and the respective trained codes; and e) naïve ratings of relationship quality to the entire set of trained codes (i.e., negativity, positivity, withdrawal, and problem solving) explained similar amounts of variance in concurrent relationship satisfaction (ps > .10). In models predicting divorce at 5-year follow-up we found a greater number of differences in fit between models with naïve ratings and models with trained codes. Although negative reciprocity and negativity, as well as HD/WW in naïve and trained codes, similarly predicted subsequent divorce, a number of other comparisons revealed differences in the predictive power of naïve versus trained ratings. Specifically, we found that models with naïve ratings had superior fit than models with trained codes in the comparison of positive reciprocity versus positivity, as well as in the comparison of rated relationship quality versus all trained ratings (differences in AICs > 2). In two instances the converse was true and trained codes had superior fit values relative to models with naïve ratings: withdrawal versus mutual avoidance and WD/HW in trained versus naïve ratings (differences in AICs > 2).

In comparing nested models (see Table 7) we found – in the majority of cases – naïve ratings and trained codes each added to the prediction of relationship satisfaction and subsequent divorce beyond the other, although there were some exceptions. Naïve ratings of negative reciprocity added to the variance in relationship satisfaction ($p < .001$) and divorce ($p < .05$) explained by the trained codes of negativity, but negativity did not account for additional variance in these outcomes when added to models including negative reciprocity. Similarly, naïve ratings of WD/HW accounted for additional variance in self-reported relationship satisfaction ($p < .05$) when added to models including trained codes of this pattern, but trained codes of WD/HW did not add to variance explained in self-reported relationship satisfaction when added to naïve ratings. Another noteworthy finding was that rated relationship quality accounted for significant additional variance in divorce when added to the full set of trained observational variables ($p < .05$) but the set of trained observational variables accounted for only a marginally significant amount of additional

variance in stability when added to rated relationship quality ($p < .10$). In one case naïve ratings did not add to significant effects of trained codes: withdrawal accounted for additional variance in relationship satisfaction when added to naive ratings of mutual avoidance ($p < .05$), but mutual avoidance did not account for significant additional variance in self-reported relationship satisfaction when added to withdrawal. Neither naive ratings nor trained codes of demand/withdraw in either direction were associated with divorce at 5-year follow-up in initial or added blocks of predictors.

Discussion

Our results support previous findings that naïve raters have an intuitive sense of communication in relationships and general relationship well-being. We found that a group of well-established dyadic interaction patterns were reliably rated by naïve observers and that these ratings were strongly and consistently associated with self-reported relationship satisfaction and trained codes both concurrently and over time in a large sample of distressed couples. Not only are these ratings linked with other important relationship outcomes, but they provide unique additional – and in a few cases superior – information about relationship functioning. Below we discuss key findings from this study in the context of previous empirical and theoretical work with couples.

Hypothesis 1 results replicated those of Waldinger et al. (2004) in that naïve ratings were strongly associated with concurrent trained codes of communication and relationship satisfaction. Our measurement of dyadic interaction patterns suggests that naïve raters can provide useful information about couples' functioning on dimensions that have well-documented links with relationship satisfaction and stability. The examination of communication over three time points allowed us to extend these and other previous findings by considering changes in communication. Consistent with Hypothesis 2, changes in naïve ratings coincided with changes in both trained codes and changes in self-reported relationship satisfaction over the course of therapy and follow-up. These findings suggest that naïve observers are sensitive to subtle changes in communication, a finding particularly important to treatment outcome research with couples where communication change is considered both an outcome and a potential mechanism for broader changes in relationship functioning. Of note, we found greater support for hypothesized links between relationship satisfaction and naïve ratings in the problem solving interactions than in the social support interactions. This pattern is consistent with the well-documented finding that negative events (e.g., communication during discussion of a major relationship problem) are more impactful to overall relationship adjustment than positive events (e.g., communication during discussion of changes one partner wants to make) (Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001), but may also reflect the greater number of problem solving interactions obtained from couples relative to social support interactions.

The most compelling finding for the use of naive raters was the similar – and at times superior – predictive ability of naïve ratings compared with trained ratings in Hypothesis 3. Not only did naïve ratings typically account for similar amounts of variance in relationship outcomes as trained codes, they were at times better predictors of these outcomes. In our comparison of independent (non-nested) models, naive ratings of negative reciprocity were superior to trained codes of negativity in predicting concurrent relationship satisfaction. Similarly, naïve ratings of positive reciprocity were superior to positivity, and naïve ratings of relationship quality were superior to the full set of trained ratings in predicting divorce at 5-year follow-up. In stepwise models our pattern of results was similar. Whereas negative reciprocity accounted for additional unique variance in satisfaction and divorce when added to models with negativity, negativity did not add to negative reciprocity's prediction of these outcomes. In the prediction of divorce at 5-year follow-up, naïve ratings of relationship

quality added to the entire set of trained codes, but trained codes accounted for only marginally significant additional variance beyond that accounted for by relationship quality.

Taken together, these findings suggest that naïve raters pick up on aspects of communication that are relevant to, and closely linked with, relationship functioning without the intensive training of traditional observational systems. This finding is important for researchers, many of whom spend a great deal of time and money obtaining ratings of couples' communication using particular systems and highly trained raters. Naïve raters open a new door for researchers: new variables can be more easily examined without creating new coding systems. Although we think this study offers a persuasive argument for the utility of naïve raters, we do not conclude that they should replace trained coders in research on communication in couples across the board. Although naïve raters seem to know *straightforward* communication when they see it, psychometrically sound assessment of less intuitive communication likely requires additional training. Naïve raters are not indicated when researchers desire information on subtleties of communication or complex behaviors, particularly those raters may not have had exposure to in their own lives (e.g., coercion; Ickes & Simpson, 1997).

There are several limitations to the current study that are important to note. One is that the naïve ratings are of interaction patterns, whereas the trained codes are couple-averages of independent codes. Partners' behavior does not exist in a vacuum; a defining characteristic of a relationship is that individuals within it are interdependent, each influenced by the other (Thibaut & Kelley, 1959). Although few would argue with the importance of dyadic interaction patterns (Weiss & Heyman, 1997), there are significant limitations to their assessment using traditional systems. Microanalytic coding systems rarely have base rates of particular behaviors high enough to be used for sequential analysis within a given couple, leading researchers to demonstrate patterns of interaction across rather than within couples (e.g., Margolin & Wampold, 1981). Aside from the Interactional Dimensions Coding System (Kline, Julien, Baucom, Hartman, Gilbert, et al., 2004), macroanalytic systems combine ratings of individual partners' behavior without establishing the temporal sequence of behaviors (e.g., in the demand/withdraw codes in this paper: ratings of wife's demanding and husband's withdrawing are summed to create the amount of WD/HW in a given interaction). Understandably, one might argue that it is our examination of these dyadic interaction patterns rather than the intuitive judgments per se that is driving the effects in this study. Although we acknowledge the unique information we have gathered with each type of system, our finding that a single item rating of relationship quality accounted for a similar amount of variance in concurrent relationship satisfaction and was a *better* predictor of subsequent divorce when compared with four highly trained communication scales (scales which are empirically-derived and made up of 3–6 items) suggests that it is not just the examination of interaction patterns relative to couple-average independent codes, but specifically intuitive judgments that offer unique information about relationships when compared with trained codes.

Another limitation is that we did not have reliable naïve ratings of all constructs included in the trained codes. Mutual avoidance scores had quite low reliability and therefore we interpret links between them and other variables with caution. Additionally, we did not have a naïve rating of problem solving or a similar construct. Although our findings with the naïve ratings included in this study are compelling, our reliability data support the notion that naïve raters are more adept at rating some interaction patterns than others. Finally, we tested the utility of undergraduate naïve raters in the assessment of communication in chronically distressed couples; it is possible the degree of rater naïveté or the extent of relationship distress could impact results of studies using naïve raters.

Nonetheless, we think our results provide a solid basis for additional observational research with couples. A natural next step would be the replication of this and other studies of naïve ratings of couples in a sample of clinicians. In clinical practice most mental health professionals do not have extensive training in observational methods with couples – or the resources and desire to get it given previously outlined limitations. We think our findings warrant research into whether clinicians can make meaningful judgments of relationship functioning and identify interaction patterns in couples without undergoing training in observational methods. Future research examining therapists' ability to rate their own clients on these dimensions will help establish the generalizability of our findings to those in the role of improving the relationships of distressed couples.

Acknowledgments

This research was supported by research grants to Andrew Christensen at UCLA (MH56223) and Neil S. Jacobson at the University of Washington (MH56165) for a two-site clinical trial of couple therapy. After Jacobson's death, William George served as PI at the University of Washington. The preparation of this manuscript was also supported by a fellowship awarded to Katherine J.W. Baucom (F31HD062168) from the Eunice Kennedy Shriver National Institute of Child Health & Human Development. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

We would like to thank Benjamin Karney, Theodore Robles, and Meghan Sweeney for their comments on an earlier draft of this manuscript.

References

Baker JK, Haltigan JD, Brewster R, Jaccard J, Messinger D. Non-expert ratings of infant and parent emotion: Concordance with expert coding and relevance to early autism risk. International Journal of Behavioral Development. 2010; 34:88–95. [PubMed: 20436947]

Baucom BR, McFarland PT, Christensen A. Gender, topic, and time in observed demand-withdraw interaction in cross- and same-sex couples. Journal of Family Psychology. 2010; 24:233–242. [PubMed: 20545396]

Baumeister RF, Bratslavsky E, Finkenauer C, Vohs KD. Bad is stronger than good. Review of General Psychology. 2001; 5:323–370.

Burnham, KP.; Anderson, DR. Model selection and multimodel inference: A practical information-theoretical approach. 2nd edition. New York: Springer; 2002.

Christensen A, Eldridge K, Catta-Preta AB, Lim VR, Santagata R. Cross-cultural consistency of the demand/withdraw interaction pattern. Journal of Marriage and Family. 2006; 68:1029–1044.

Coan, JA.; Gottman, JM. The Specific Affect (SPAFF) coding system. In: Coan, JA.; Allen, JJB., editors. Handbook of emotion elicitation and assessment. New York, NY: Oxford University Press; 2007. p. 106-123.

Cohen, J.; Cohen, P.; West, SG.; Aiken, LS. Applied multiple regression/correlation analysis for the behavioral sciences, 3rd. Mahwah, NJ: Lawrence Erlbaum Associates; 2003.

Ebling R, Levenson RW. Who are the marital experts? Journal of Marriage and Family. 2003; 65:130–142.

Eldridge KA, Sevier M, Jones J, Atkins DC, Christensen A. Demand-withdraw communication in severely distressed, moderately distressed, and nondistressed couples: Rigidity and polarity during relationship and personal problem discussions. Journal of Family Psychology. 2007; 21:218–226. [PubMed: 17605544]

Gottman, JM. The relationship between marital processes and marital outcomes. Hillsdale, NJ, England: Lawrence Erlbaum Associates; 1994. What predicts divorce?.

Gottman, JM.; McCoy, K.; Coan, J.; Collier, H. The specific affect coding system (SPAFF) for observing emotional communication in marital and family interaction. Mahwah, NJ: Erlbaum; 1995.

Heyman RE. Observation of couple conflicts: Clinical assessment applications, stubborn truths, and shaky foundations. Psychological Assessment. 2001; 13:5–35. [PubMed: 11281039]

Heyman, RE. Rapid Marital Interaction Coding System. In: Kerig, PK.; Baucom, DH., editors. Couple observational coding systems. Mahwah, NJ: Erlbaum; 2004. p. 67-94.

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

Ickes, W.; Simpson, JA. Managing empathic accuracy in close relationships. In: Ickes, W., editor. Empathic accuracy. New York: Guilford; 1997. p. 218-250.

Kenny, DA.; Kashy, DA.; Cook, WL. Dyadic data analysis. New York: Guilford; 2006.

Kerig, PK.; Baucom, DH. Couple observational coding systems. NJ: Erlbaum; 2004.

Kline, GH.; Julien, D.; Baucom, B.; Hartman, S.; Gilbert, K.; Gonzales, T., et al. The Interactional Dimensions Coding System: A global system for couple interactions. In: Kerig, P.; Baucom, D., editors. Couple observational coding systems. Mahwah, NJ: Erlbaum; 2004.

Lorber MF. Can minimally trained observers provide valid global ratings? Journal of Family Psychology. 2006; 20:335–338. [PubMed: 16756410]

Margolin G, Wampold BE. Sequential analysis of conflict and accord in distressed and nondistressed marital partners. Journal of Consulting and Clinical Psychology. 1981; 49:554–567. [PubMed: 7264037]

Noller P, White A. The validity of the Communication Patterns Questionnaire. Psychological Assessment: A Journal of Consulting and Clinical Psychology. 1990; 2:478–482.

Snyder DK, Heyman RE, Haynes SN. Evidence-based approaches to assessing couple distress. Psychological Assessment. 2005; 17:288–307. [PubMed: 16262455]

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.

StataCorp. Stata Statistical Software: Release 11. College Station, TX: StataCorp LP; 2009.

Raudenbush, SW.; Bryk, AS.; Cheong, YF.; Congdon, R. Lincolnwood, IL: Scientific Software International; 2007. HLM 6.04: Hierarchical Linear and Nonlinear Modeling.

Thibaut, JW.; Kelley, HH. The social psychology of groups. New York: Wiley; 1959.

Vuong QH. Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica. 1989; 57:307–333.

Waldinger RJ, Schulz MS, Hauser ST, Allen JP, Crowell JA. Reading others' emotions: The role of intuitive judgments in predicting marital satisfaction, quality, and stability. Journal of Family Psychology. 2004; 18:58–71. [PubMed: 14992610]

Weiss, RL.; Heyman, RE. A clinical-research overview of couples interaction. In: Halford, WK.; Markman, HJ., editors. Clinical handbook of marriage and couples intervention. NY: Wiley & Sons; 1997. p. 13-41.

Weiss, RL.; Heyman, RE. Couples observational research: An impertinent, critical overview. In: Kerig, PK.; Baucom, DH., editors. Couple observational coding systems. NJ: Erlbaum; 2004. p. 11-26.

Table 1

Descriptive Statistics, Reliabilities, and Partial Correlations between Naïve Codes

Code	*M*	*SD*	α	1	2	3	4	5
1. Quality	50.57	11.18	.80					
2. Negative Reciprocity	3.31	1.81	.80	−.76***				
3. Positive Reciprocity	3.26	1.45	.70	.80***	−.68***			
4. WDHW	3.19	1.49	.63	−.41***	.45***	−.42***		
5. HDWW	2.80	1.18	.64	−.38***	.45***	−.33***	−.12***	
6. Mutual avoidance	2.23	0.93	.43	−.24***	.20***	−.17***	.24***	.17***

Note. α = interobserver reliability. All means and standard deviations are across time (pre-therapy, post-therapy, 2-year follow-up), type of discussion (problem solving, social support), and who chose the topic (husband, wife). Interobserver Cronbach's alphas were computed separately within each topic of each type of discussion at each time point and averaged. All correlations are controlling for time, type of discussion, and who chose the topic.

* $p < .05$,

** $p < .01$,

*** $p < .001$,

$N = 134$

Table 2

Concurrent Associations between Naïve Ratings and Trained Ratings

	Negativity		Positivity		Withdrawal		Problem Solving	
	B (SE)	***SC***	***B (SE)***	***SC***	***B (SE)***	***SC***	***B (SE)***	***SC***
Quality	−0.05 (0.00)***	−0.45	0.04 (0.00)***	0.46	−0.02 (0.00)***	−0.30	0.04 (0.00)***	0.39
Negative Reciprocity	0.53 (0.02)***	0.74	−0.24 (0.02)***	−0.48	0.09 (0.01)***	0.26	−0.25 (0.02)***	−0.44
Positive Reciprocity	−0.33 (0.03)***	−0.37	0.33 (0.02)***	0.54	−0.09 (0.02)***	−0.21	0.26 (0.03)***	0.37
WDHW	0.23 (0.04)***	0.26	−0.14 (0.02)***	−0.24	0.09 (0.02)***	0.21	−0.13 (0.03)***	−0.18
HDWW	0.36 (0.03)***	0.33	−0.13 (0.02)***	−0.17	0.12 (0.02)***	0.21	−0.15 (0.03)***	−0.17
Mutual avoidance	−0.04 (0.03)	−0.03	−0.02 (0.03)	−0.02	0.28 (0.03)***	0.39	−0.04 (0.04)	−0.04

Note. *SC* = standardized regression coefficient (calculated as unstandardized coefficient times SD of predictor over SD of trained rating), Unstandardized regression coefficients for effects of naïve codes on couple-level (i.e., husband and wife average) relationship outcomes controlling for time (pre-therapy, post-therapy, 2-year follow-up), type of discussion (problem solving, social support), who chose the topic (husband, wife), and differences between husband and wife in the outcome variable.

$^{\dagger}p < .10$,

$^{*}p < .05$,

$^{**}p < .01$,

$^{***}p < .001$

Table 3

Concurrent Associations between Naïve Ratings in Problem Solving and Social Support Discussions and Relationship Satisfaction

	Problem Solving		Social Support	
	B (SE)	***SC***	***B (SE)***	***SC***
Quality	0.44 (0.06)***	0.37	0.12 (0.06)†	0.05
Negative Reciprocity	−2.89 (.35)***	−0.33	−0.28 (0.49)	−0.02
Positive Reciprocity	2.28 (0.33)***	0.25	0.75 (0.46)	0.05
WDHW	−1.19 (0.31)***	−0.12	−1.01 (0.69)	−0.05
HDWW	−1.42 (0.37)***	−0.12	−0.95 (0.53)†	−0.06
Mutual Avoidance	0.12 (0.34)	0.01	−1.67 (0.63)**	−0.09

Note. *SC* = standardized regression coefficient (calculated as unstandardized coefficient times SD of predictor over SD of relationship satisfaction). Unstandardized regression coefficients for unique effects of naïve codes from each type of discussion on couple-level relationship satisfaction; controlling for time (pre-therapy, post-therapy, 2-year follow-up) and difference between husband and wife relationship satisfaction.

† $p < .10$,

* $p < .05$,

** $p < .01$,

*** $p < .001$

Table 4

Associations between Changes in Naïve Ratings and Changes in Trained Ratings

	Negativity		Positivity		Withdrawal		Problem Solving	
	B (SE)	η^2	***B (SE)***	η^2	***B (SE)***	η^2	***B (SE)***	η^2
Quality	−0.06 (0.01)***	.36	0.05 (0.00)***	.69	−0.02 (0.00)***	.17	0.05 (0.01)***	.61
Neg Reciprocity	0.52 (0.03)***	.58	−0.27 (0.03)***	.58	0.13 (0.02)***	.19	−0.29 (0.03)***	.60
Pos Reciprocity	−0.31 (0.03)***	.44	0.35 (0.03)***	.85	−0.11 (0.02)***	.17	0.29 (0.03)***	.51
WDHW	0.32 (0.04)***	.26	−0.18 (0.03)***	.21	0.09 (0.03)***	.06	−0.16 (0.04)***	.12
HDWW	0.30 (0.04)***	.24	−0.19 (0.03)***	.25	0.16 (0.03)***	.18	−0.21 (0.05)***	.15
Mutual avoidance	−0.01 (0.04)	.00	−0.06 (0.05)	.01	0.32 (0.04)***	.32	−0.10 (0.05)†	.02

Note. η^2= eta squared as computed with Stata program. Unstandardized regression coefficients can be interpreted as associations between changes in naïve ratings and changes in trained ratings since we controlled for previous timepoint's naïve and trained ratings (Cohen, Cohen, West, & Aiken, 2003). We also controlled the timepoint predicted (post-therapy, 2-year follow-up), type of discussion (problem solving, social support), and who chose the topic (husband, wife). All variables were centered.

† $p < .10$,

* $p < .05$,

** $p < .01$,

*** $p < .001$

Table 5

Associations between Changes in Naïve Ratings and Changes in Relationship Satisfaction

	Problem Solving		Social Support	
	B (SE)	η^2	***B (SE)***	η^2
Quality	0.67 (0.08)***	.35	0.52 (0.14)***	.07
Negative Reciprocity	−4.12 (0.53)***	.35	−0.54 (1.06)	.00
Positive Reciprocity	3.40 (0.45)***	.33	3.23 (0.91)***	.07
WDHW	−2.12 (0.60)***	.09	−1.89 (1.70)	.01
HDWW	−1.81 (0.75)*	.04	−1.81 (1.02)†	.02
Mutual Avoidance	0.29 (0.90)	.00	−3.88 (1.35)**	.06

Note. η^2= eta squared as computed with Stata program. Unstandardized regression coefficients for unique effects of changes in naïve codes from each type of discussion on changes in couple-level relationship satisfaction (i.e., controlling both naïve and trained ratings from previous time point).

† $p < .10$,

* $p < .05$,

** $p < .01$,

*** $p < .001$.

Table 6

Prediction of Relationship Satisfaction and Long-Term Status by Naïve versus Trained Ratings: Comparison of Non-nested Models

	Concurrent Relationship Satisfaction			
Variables in Each Model	***R*²**	**Vuong's *z***		
Negative reciprocity	0.29			
Negativity	0.24	2.50*		
Positive reciprocity	0.24			
Positivity	0.24	0.25		
Mutual avoidance	0.13			
Withdrawal	0.15	−1.07		
HDWW (naïve)	0.17			
HDWW (trained)	0.17	0.25		
WDHW (naïve)	0.17			
WDHW (trained)	0.15	1.12		
Quality	0.29			
All trained ratings	0.33	−1.46		
	Divorce at 5-year Follow-Up			
Variables in Each Model	Pseudo-R^2	*AIC*	diff*AIC*	
Negative reciprocity	0.078	499.04		
Negativity	0.076	499.86	−0.82	
Positive reciprocity	0.061	507.62		
Positivity	0.035	521.73	−14.11	+
Mutual avoidance	0.036	521.12		
Withdrawal	0.053	512.31	8.81	+
HDWW (naïve)	0.040	518.83		
HDWW (trained)	0.043	517.63	1.20	
WDHW (naïve)	0.040	519.01		
WDHW (trained)	0.055	511.28	7.73	+
Quality	0.096	489.31		
All trained ratings	0.104	497.00	−7.69	+

Note. *AIC* = Akaike's Information Criterion, All trained ratings = negativity, positivity, withdrawal, and problem solving. All models included time (pre-therapy, post-therapy, 2-year follow-up) and separate predictors from each type of discussion (problem solving, social support) in the model. In models where relationship satisfaction was the outcome we also controlled for difference between husband's and wife's relationship satisfaction.

† $p < .10$,

* $p < .05$,

** $p < .01$,

*** $p < .001$,

+indicates difference in AIC greater than 2 (negative diff*AIC*values suggest models with naïve codes have better fit relative to models with trained codes since *smaller* AICs indicate greater model fit).

Table 7

Prediction of Relationship Satisfaction and Long-Term Status by Naïve versus Trained Ratings: Comparison of Nested Models

	Concurrent Relationship Satisfaction		
	Initial block R^2		
Block 1[a]: Time, diffDAS	0.1122		
	Variables added to initial block		
Block 2 variables	ΔR^2	Block 3 variables	ΔR^2
Negativity	.13***	Negative reciprocity	.05***
Negative reciprocity	.18***	Negativity	.00
Positivity	.13***	Positive reciprocity	.04***
Positive reciprocity	.13***	Positivity	.03***
All trained ratings	.22***	Quality	.02*
Quality	.18***	All trained ratings	.07***
Withdrawal	.04**	Mutual avoidance	.00
Mutual avoidance	.02*	Withdrawal	.02*
HDWW (trained)	.06***	HDWW (naïve)	.02**
HDWW (naïve)	.06***	HDWW (trained)	.01*
WDHW (trained)	.04**	WDHW (naïve)	.02*
WDHW (naïve)	.06***	WDHW (trained)	.00
	Divorce at 5-year follow-up		
	Initial block χ^2		
Block 1[a]: Time	20.31***		
	Variables added to initial block		
Block 2 variables	Wald χ^2	Block 3 variables	Wald χ^2
Negativity	6.63*	Negative reciprocity	6.18*
Negative reciprocity	10.46**	Negativity	2.99
Positivity	0.84	Positive reciprocity	12.50**
Positive reciprocity	5.97†	Positivity	6.43*
All trained ratings	14.95†	Quality	7.21*
Quality	10.04**	All trained ratings	14.00†
Withdrawal	5.80†	Mutual avoidance	6.42*
Mutual avoidance	1.16	Withdrawal	10.14**
HDWW (trained)	1.83	HDWW (naïve)	0.33

HDWW (naïve)	1.36	HDWW (trained)	1.06
WDHW (trained)	4.05	WDHW (naïve)	0.03
WDHW (naïve)	1.82	WDHW (trained)	2.87

Note. diffDAS = difference between husband's and wife's relationship satisfaction, All trained ratings = negativity, positivity, withdrawal, and problem solving. Each row represents a separate model. All models included separate predictors from each type of discussion (problem solving, social support) in the model.

[a] Initial block for all models that follow.

† $p < .10$,

* $p < .05$,

** $p < .01$,

*** $p < .001$