Journal of Family Theory & Review

FRANK D. FINCHAM *Florida State University*

RON ROGGE *University of Rochester**

Understanding Relationship Quality: Theoretical Challenges and New Tools for Assessment

Relationship quality is studied in a variety of disciplines, yet widely accepted practices promulgate a lack of conceptual clarity. We build on a conceptually simple and theoretically advantageous view of relationship quality and suggest a shift to conceptualizing it as two distinct yet related dimensions—positive and negative evaluations of relationships. We introduce item response theory as a powerful tool for measure development, demonstrating how relationship quality can be optimally pursued in the context of modern test theory, thus leading to better theory development. Recognizing the limitations of self-reported relationship quality, we extend this two-dimensional conceptualization further by drawing on developments in the derivation of implicit measures. After briefly introducing such measures, we illustrate their application to assessment of relationship quality.

Since its inception, marital research has been strongly motivated by the desire to understand and remediate family problems. Because subjective global evaluation of the relationship (relationship quality) is viewed as the final common pathway that leads to relationship breakdown (Jacobson, 1985), it has been the dominant construct studied in the literature on relationships such as marriage. Not surprisingly, it has gained the attention of researchers from a variety of disciplines, including psychology, sociology, family studies, and communication.

Unfortunately, the wealth of empirical attention given to the study of marriage is inversely proportional to conceptual analysis of the central construct studied. As Glenn (1990) pointed out in his decade review of research on marriage, most studies are justified on practical grounds, ''with elements of theory being brought in on an incidental, ad hoc basis'' (p. 818). The enemy of scientific progress, conceptual confusion, has resulted and the literature is littered with a large number of terms, such as satisfaction, adjustment, success, happiness, companionship, or some synonym reflective of the quality of the relationship. These terms tend to be used interchangeably leading some scholars to even call for the elimination of such terms as *marital satisfaction* and *marital adjustment* from the literature (e.g., Trost, 1985). Rather than heeding such a call (and thereby trying to avoid the problem), we confront the conceptual challenge directly in this article. This article focuses on a partner's subjective evaluation of a romantic relationship, and hence we prefer the term *relationship quality*. However, because such judgments have also been referred to as relationship satisfaction (Funk & Rogge, 2007), we use the terms *relationship quality* and *relationship satisfaction* interchangeably.

Family Institute, 225 Sandels Building, Florida State University, Tallahassee, FL 32306-1491 (ffincham@fsu.edu).

*Department of Clinical and Social Psychology, University of Rochester, RC Box 270266, Rochester, NY 14627-0266 (ronald.rogge@rochester.edu).

Key Words: marital satisfaction, relationship quality.

In the first section, we provide a brief synopsis of current concerns regarding the construct of relationship quality, including the psychometric limitations of current measures. This serves as a springboard for introducing item response theory and demonstrating how it can advance measurement of relationship quality.[1] We then suggest a fundamental shift in how relationship quality is conceptualized, arguing that it might be more appropriately represented by two distinct but related dimensions—positive and negative relationship qualities. Given the limitations of self-report measures, in the final section, we describe how implicit measures might be used to evaluate positive and negative dimensions of relationship quality, offering insights into relationship functioning of which subjects might be unaware or unwilling to report. The article concludes by summarizing the critique offered and by highlighting avenues for future research.

Will the Real Relationship Satisfaction, Adjustment, Success, Happiness, and Quality Please Stand Up?

As indicated by the title of this section, numerous terms have been used to refer to what we earlier identified as the final common pathway to relationship breakdown. Terminological heterogeneity combined with disciplinary diversity means that relevant material is ''scattered across a variety of disparate sources,'' which makes it extremely difficult ''to access the picture of marriage painted by scientific research'' (Fincham, 1998, p. 543).

Nonetheless, there is increasing recognition of two major approaches to the central construct studied by marital researchers. They focus on the relationship and on intrapersonal processes, respectively. The relationship or interpersonal approach typically looks at patterns of interaction such as companionship, conflict, and communication and tends to favor use of such terms as *adjustment*. In contrast, the intrapersonal approach focuses on individual judgments of spouses, namely their subjective evaluation of the marriage. This approach tends to use such terms as *marital satisfaction* and *marital happiness*. Both approaches are valuable ones, but problems arise because of another feature of research on relationship quality, to which we now turn.

Relationship quality and relationship satisfaction (and similar constructs denoted by various synonyms) have been almost exclusively assessed using self-report. Ironically, even behaviorally oriented psychologists who rejected the utility of self-report when they began to study marriage systematically in the 1970s used self-reported satisfaction as a criterion variable in their studies (see Bradbury & Fincham, 1987). Indeed, a primary goal was to account for variability in such reports of relationship quality using coded observations of marital interactions. The problem is that the most widely used measures of relationship quality, the Marital Adjustment Test (MAT) (Locke & Wallace, 1959) and the Dyadic Adjustment Scale (DAS) (Spanier, 1976), include items that assess both interaction patterns (interpersonal processes) and subjective evaluations of the marriage (intrapersonal processes), thereby ignoring the conceptual distinction just outlined. Historically, it has been the case, then, that researchers study the same thing regardless of conceptual distinctions they might make when introducing their research.

It is also questionable whether spouses are the best, or even good, reporters of relationship properties. Indeed, research using the Spouse Observation Checklist shows that, on average, spouses agree only about 50% of the time on the occurrence of a behavior, and as a result, the epistemological status of such reports has changed so that they are now viewed not as objective reports of behavior (as originally thought) but as subjective perceptions (for a review, see Bradbury & Fincham, 1987). This viewpoint is consistent with the construct of sentiment override (Weiss, 1980) whereby spouses are posited to respond to questionnaire items about the partner and marriage, not so much in terms of the item's manifest content but in terms of their sentiment toward the partner. Self-report seems better suited to the second major approach to marital quality, which focuses on how spouses evaluate their marriage.[2]

[1] Although most of our observations are motivated by characteristics of marital research, we use the term *relationship quality* wherever appropriate because it is more inclusive.

[2] Clearly, some properties of the relationship can be obtained only from spouses (e.g., frequency of intercourse), but others may be beyond the awareness of all but the most psychologically sophisticated (e.g., the pattern of interaction during conflict).

In light of the foregoing observations, it is perhaps little wonder that, early on, factor analytic approaches gave rise to the conclusion that ''different operations designed to measure marital satisfaction converge and form one dimension'' (Gottman, 1979, p. 5), a viewpoint supported by subsequent work that shows that standard measures of relationship satisfaction intercorrelate highly (e.g., Heyman, Sayers, & Bellack, 1994). But what are we to make of relationship quality assessed in this manner? Dahlstrom (1969) describes three levels at which responses to self-report inventories can be interpreted: (a) veridical descriptions of behavior (e.g., responses regarding frequency of disagreement reflect the actual rate of disagreement between spouses); (b) potential reflections of attitudes (e.g., frequently reported disagreement may reflect high rates of disagreement but may also reflect the view that the partner is unreasonable, that the spouse feels undervalued, or some other attitude); and (c) behavioral signs, the meaning of which can be determined only by actuarial data (e.g., rated disagreement may reflect time spent together, respondents' self-esteem, frequency of intercourse, or a host of other variables). Few measures of relationship quality specify the level at which responses are to be interpreted.

The summation of various dimensions of relationships in omnibus measures of relationship quality (e.g., interaction, happiness) also precludes meaningful study of the interplay of such dimensions (e.g., interaction may influence satisfaction, and vice versa). It has also given rise to another troubling feature of the literature on relationship quality, namely that our knowledge of the determinants and correlates of relationship quality includes (an unknown number of) spurious findings. This is because of overlapping item content in measures of quality and measures of constructs examined in relation to it. The often documented association between self-reported communication (e.g., Marital Communication Inventory; ''Do the two of you argue a lot over money?'' ''Do you and your and your spouse engage in outside activities together?'') and relationship quality (e.g., DAS; ''Indicate the extent of agreement or disagreement between you and your partner on: handling family finances''; ''Do you and your mate engage in outside interests together?'') is a particularly egregious example of this problem. The resulting tautological association hinders theory construction and affects the credibility of research findings. Funk and Rogge (2007) offered evidence to support the cross-contamination of communication and relationship quality measures, identifying 13 items from the MAT and DAS that correlated more strongly with the communication factor (extracted from a principle components analysis of 176 satisfaction and communication items in a sample of 5,315 respondents) than they did with the satisfaction factor. Fincham and Bradbury (1987) discuss the dilemma caused by overlapping item content at some length, showing that exclusion of the items common to both measures does not provide a satisfactory solution to this problem, as they usually reflect overlap in the definition of the constructs.

Perhaps the most important issue confronting the field is that the link between conceptual distinctions and measurements is not as strong as it might be, thereby hindering theory development. But this is hardly surprising when, in the broader marital literature, ''the association between theories and research tends to be loose and imprecise and, in some cases, constitutes only a metaphorical connection'' (Fincham & Beach, 1999, p. 55). As might be expected, this has an impact at the level of assessment. It is therefore important for assumptions underlying the measurement of relationship quality to be made explicit and questioned. For example, Spanier (1976) eliminated items from his influential measure (the DAS) that were positively skewed, thereby assuming that items reflective of relationship quality approximate a normal distribution. But as Norton (1983) pointed out, such items may be less critical indicators or even irrelevant to relationship quality if relationship quality inherently involves skewed data because spouses tend to report happy marriages. Moreover, if the outcome predicted by relationship quality is itself skewed (e.g., aggression), then a skewed predictor may be best (Heyman et al., 1994).

Some time ago, a leading scholar concluded that the ''psychometric foundation is reasonably solid and need not be redone'' (Gottman & Levenson, 1984, p. 71). The basis for such a conclusion appears to be that different measures of relationship quality intercorrelate highly, thus suggesting that differences in item content across measures are relatively unimportant (e.g., Funk & Rogge, 2007; Heyman et al., 1994).

Such conclusions are quite reasonable for some research purposes. For instance, they

suffice if the goal is to select happy or satisfied versus unhappy or dissatisfied spouses, as is often done in clinical research on marriage. Here the exact content of the measure used to select groups is less important than its ability to identify correctly the groups of interest. In fact, it was precisely this type of criterion keying (identifying items optimally distinguishing between distressed and nondistressed couples) that was used to create the two most widely used and cited measures of relationship satisfaction: the MAT and DAS scales. From an actuarial point of view, such a taxometric approach has merit, as it can effectively create a tool to identify risk groups. However, to the extent that one's goal is to develop theory for advancing understanding of relationship quality as a continuous construct or to devise conceptually sound measures of relationship quality across all ranges of functioning, the above approach is less appropriate.

In light of the foregoing observations, we argue that current conceptions and operationalizations of relationship quality are inadequate. Much of the conceptual confusion regarding relationship quality appears to be based on the assumption that constructs related at the empirical level are equivalent at the conceptual level. This can lead to a problem that is demonstrated by considering the example of height and weight. Those two dimensions correlate to about the same degree as many measures of relationship quality, yet much is gained by keeping height and weight separate. Imagine designing a door frame having only a composite measure of the bigness of users and not their height! Keeping empirical and conceptual levels of analyses separate has the advantage of forcing the researcher to articulate the nature of the construct and the domain of observables to which it relates before developing measures of the construct. Such practices are likely to facilitate theoretical development and the construction of more easily interpreted measures of relationship quality.

Mindful of the conceptual problems encountered in using omnibus measures of relationship quality, we build on previous work that has defined relationship quality as subjective, global evaluation of the relationship (e.g., Fincham & Bradbury, 1987; Norton, 1983; Schumm, Nichols, Schectman, & Grinsby, 1983). The strength of this approach is its conceptual simplicity, as it avoids the problem of interpretation and allows for unambiguous specification of the construct's nomological network. Because it has a clear-cut interpretation, the approach allows the antecedents, correlates, and consequences of relationship quality to be examined in a straightforward manner.

The remainder of the article builds on this prior work in three ways. First, with one exception, relationship satisfaction indices have been developed using classical test theory and therefore fail to take advantage of recent developments in psychometrics. This is important because, as Funk and Rogge (2007) have noted, extant scales of relationship quality have problematic levels of noise in measurement, reducing their power to reveal meaningful results and thereby hindering theory development. The next section therefore introduces a more recent development in psychometrics, item response theory (IRT), also known as latent trait theory, and illustrates its application to the assessment of relationship quality. After doing this, we suggest a fundamental shift in how relationship quality is conceptualized, arguing that it might be more appropriately represented by two distinct but related dimensions—positive and negative relationship qualities. This mirrors robust findings in the affect literature exemplified by scales like the Positive Affect Negative Affect Schedule (PANAS) (Watson, Clark, & Tellegen, 1988) and the Mood and Anxiety Symptom Questionnaire (MASQ) (Watson et al., 1995), suggesting that the experience of positive and negative affect or distress and vitality are substantively distinct phenomena. We argue that individuals in romantic relationships might similarly experience both positive and negative feelings toward their relationships somewhat independently, and therefore constraining the assessment of relationship quality to a single dimension could obscure results and oversimplify theories. In the fourth section, we turn to consider the limitations of self-report measures. Although the bulk of relationship research has made use of self-report scales to assess relationship quality, that method of assessment is, by definition, limited by subjects' own awareness of and insight into their true level of relationship quality. Self-report measures have also been shown to be affected by a number of common biases, such as impression management and motivated distortion (e.g., Stone et al., 2000). Thus, in the fourth section, we examine how positive and negative relationship evaluations might be explored using implicit measures, thus offering the possibility of accessing information on relationship quality

of which subjects might not be fully aware or willing to fully report. Such information has the potential to enrich theory.

Placing Relationship Quality in the Context of Modern Test Theory: Enter Item Response Theory

To explain how item response theory (Hambleton, Swaminathan, & Rogers, 1991) can shed light on the assessment of relationship quality, and thereby help promote theory development, we first need to provide a brief description of IRT. The field of standardized testing has long used IRT to craft nonidentical but equivalent forms of tests evaluating academic ability and competency. Limited primarily by the exceedingly large sample sizes required, and to a lesser extent by the complexity of calculations involved, IRT offers several important advantages over classical test theory approaches.

One advantage is that when an item is evaluated with IRT in a sufficiently large and diverse sample, the results obtained can be expected to replicate almost identically in all future samples. This provides insight into how that item will perform across a range of situations and clarifies exactly how much information it will provide for assessing the construct of interest (θ in IRT). This also means that a score for a measure developed using IRT should have an identical meaning across samples. A second advantage offered by IRT is that it assumes that the utility of individual items varies by levels of the construct being assessed (θ). Put simply, this means that some items might be highly effective at assessing low levels (e.g., -2 *SD* to -1 *SD* below the population mean) of a construct like relationship satisfaction. Conversely, other items might offer little information for people in that range but could offer large amounts of information for assessing higher levels (e.g., $+1$ *SD* to $+2$ *SD*) of that same construct. Although the possibility of such differences has long been recognized in the measurement literature, classical test theory techniques like factor loadings, item-to-total correlations, squared multiple correlations, and Cronbach's alpha coefficients are simply unable to reveal such differences or to quantify them as clearly as IRT. Finally, IRT is able to synthesize the results into information profiles for each item (called item information curves, or IICs) that reveal how much information an item provides for assessing the construct of interest at various levels of that construct. In this framework, the standard error of measurement is the inverse square root of the information curve. Thus, by quantifying the information provided by each item at various levels of θ, IRT also quantifies the noise in measurement at various levels of θ. As a result, IRT offers a powerful technique for evaluating the precision of measurement afforded by individual items as well as sets of items (or scales).

To explain how IRT accomplishes all of this, it is necessary to first explain the basic mechanics of IRT analyses. An IRT analysis begins by estimating latent scores on the construct of interest for each individual in a sample (θ scores). These would be the equivalent of GRE or SAT scores. Then, on the basis of those θ estimates, IRT estimates a set of parameters for each item in the analysis (α and β coefficients). When evaluating each item, IRT is simply looking to see whether higher θ scores are associated with respondents' selection of higher response choices on that item. To the degree that higher θ scores are tightly linked to higher response choices so that there are very sharp boundaries between response choices, an item is considered highly informative. If, by contrast, responses on the item seem to be relatively unrelated to θ scores, then the item would be deemed to offer little information for assessing θ. After estimating item parameters for all of the items (and therefore shedding light on which items offer the most information), IRT starts another iterative cycle by reestimating θ scores for each individual in the sample on the basis of the new item parameters, giving greater weight to the items offering greater amounts of information. With those new θ scores, IRT then reestimates the item parameters. This iterative process stops when both the θ scores and the item parameters stabilize. The appendix describes this process in more detail for dichotomous and Likert scale items.

Application to Relationship Quality

In terms of assessing relationship quality, IRT offers a number of exciting possibilities. First and foremost, IRT offers the chance to quantify the precision of measurement (lack of noise) offered by current relationship quality scales. Although 40 years of converging data offer strong evidence that measures like the DAS and MAT are indeed measuring relationship

quality, very little attention has been given to determining how precisely or accurately they assess that construct. This would be equivalent to doing 40 years of research studying fever medications using the same one or two brands of thermometers without knowing whether they were accurate to ±0.1 degrees or ±10 degrees. As long as the thermometers were indeed measuring temperature, researchers should still get converging results. However, if researchers were using thermometers that were accurate to only ±10 degrees, it would take considerably larger sample sizes to discover reliable patterns of change over time and such extreme noise in measurement would likely obscure significant and meaningful results in smaller samples. An important casualty of such circumstances is likely to be theory development. In many ways, couples researchers find themselves in precisely this position, and although research using scales like the MAT and DAS has undoubtedly been fruitful, if those scales were to have notably low levels of precision (high measurement noise), then the countless significant results that excessive noise was likely to have masked would outweigh the information gained by using those measures.

To address this issue, IRT analyses were applied to the items of existing relationship quality scales (e.g., MAT, DAS, Norton's [1983] Quality of Marriage Index [QMI], Hendrick's [1988] Relationship Assessment Scale [RAS], Schumm et al.'s [1983] Kansas Marital Satisfaction Scale [KMS], and a 15-item Semantic Differential [SMD] [Karney & Bradbury, 1997]) in a sample of 5,315 online respondents (Funk & Rogge, 2007). Although the IICs suggested that a number of items from existing scales offered high levels of information for assessing relationship quality, many of the items provided notably low levels of information, indicating that responses to those items were relatively unrelated to relationship quality and would therefore primarily contribute error variance or noise to the scales that used them. This was born out in the test information curves, as the 32-item DAS seemed to offer little more information than the 6-item QMI, and the 15-item MAT offered no more information than a 4-item version of the DAS. Thus, the IRT analyses suggested that the two most widely used and cited measures of relationship quality, the MAT and DAS, had markedly high levels of measurement noise.

In addition to offering a powerful tool for evaluating current relationship quality scales, IRT also offers the possibility of developing psychometrically optimized scales. Given the high levels of noise in the MAT and DAS, Funk and Rogge (2007) created the Couples Satisfaction Index (CSI) by using a combination of exploratory factor analyses and IRT analyses on a pool of 140 items to identify the unidimensional, nonredundant set of 32, 16, and 4 items offering the greatest information (lowest noise) for assessing relationship quality. The CSI scales offered identical patterns of correlation with anchor scales from the nomological net to those obtained with the MAT and DAS and demonstrated appropriately high levels of correlation with scales like the MAT and DAS (all correlations greater than .87), which suggests that they were still assessing the same construct of relationship quality. In fact, all of the measures of relationship quality demonstrated exceedingly high levels of correlation with one another and comparable patterns of correlation with anchor scales. Thus, at a correlational level (in a large and diverse sample), the measures seemed completely interchangeable. However, Funk and Rogge (2007) were able to demonstrate that the increased information offered by the CSI sales translated into increased precision (decreased noise) and markedly higher levels of power for detecting group differences than the MAT and DAS.

The psychometric analyses generating the CSI scales also shed light on the theoretical underpinnings of the construct of relationship quality. As both the MAT and DAS were constructed primarily using criterion keying (selecting items optimally separating distress from nondistressed couples), the resultant scales had markedly heterogeneous item content, which leads to theoretical uncertainty in the boundaries of the construct. In contrast, by using statistical techniques more appropriate to evaluating and developing measures of continuous constructs, the IRT analyses presented in Funk and Rogge (2007) identified a more homogeneous set of items for the CSI scales. Specifically, items identified as most informative by the IRT analyses also happened to be the items most prototypical of the global evaluative dimension (e.g., the top four items included ''Please indicate the degree of happiness, all things considered, of your relationship,'' ''I have a warm and comfortable relationship with my partner,'' ''How rewarding

is your relationship with your partner?'' and ''In general, how satisfied are you with your relationship?''). This suggests that respondents' methods of responding to relationship quality items align most directly with the more focused theoretical definition of this construct discussed earlier. Thus, the development of the CSI scales offers relationship researchers a much better set of thermometers for evaluating relationship quality. Moreover, the reduced measurement error associated with those scales offers the possibility of having greater power to detect theoretically meaningful results—particularly in small samples. The findings offer a direct challenge to the earlier cited assertion that for measures of relationship quality the ''psychometric foundation is reasonably solid and need not be redone'' (Gottman & Levenson, 1984, p. 71). Instead, for the past 40 years, measurement noise has been a serious problem lurking underneath the seemingly robust and convergent findings of studies using the DAS and MAT. This lack of precision provides another reason that helps account for the relative lack of theoretical development in the marital literature.

Broadening Our Horizons: A Two-Dimensional Conceptualization of Relationship Quality

The bulk of couple research has assumed that relationship quality represents a single bipolar dimension ranging from extreme dissatisfaction to extreme satisfaction. However, Fincham and Linfield (1997) challenged this assumption, arguing that individuals might be able to simultaneously hold both negative and positive sentiments toward romantic partners, just as individuals can experience both positive and negative affect at the same time. Fincham and Linfield went on to hypothesize that assessing the two dimensions independently of each other would provide additional information on current relationship functioning that could not be obtained from unidimensional measures like the MAT and DAS. To test this, Fincham and Linfield developed two 3-item scales to assess each dimension, the Positive Marital Quality (PMQ) and the Negative Marital Quality (NMQ) scales. To enhance the distinction between the two dimensions, the beginning of each item asked respondents to consider only the dimension they were evaluating (e.g., ''Considering only the positive qualities of your spouse, and ignoring the negative ones, evaluate how positive these qualities are''). Using a sample of 123 married couples, the authors demonstrated that PMQ and NMQ each accounted for unique variance in self-reports of conflict behavior and attributions even after controlling for MAT scores. Thus, their results suggested that new and useful information was gained by disentangling the assessment of positive sentiment toward a relationship from negative sentiment toward that same relationship. Mattson, Paldino, and Johnson (2007) have also shown the utility of assessing positive and negative quality separately using the PMQ and NMQ among engaged couples.

Extending these results to the evaluation of treatment effects over time, Rogge et al. (2010) examined linear change in relationship quality over 3 years using the MAT, PMQ, and NMQ in a sample of 174 couples who had received either no treatment (NoTx, $n = 44$), the Preparation and Relationship Enhancement Program (PREP) (Jacobson & Margolin, 1979; $n = 45$), the Compassionate and Accepting Relationships Through Empathy program (CARE) (Rogge, Cobb, Johnson, Lawrence, & Bradbury, 2002; $n = 52$), or an intervention designed to increase couples' awareness of their own relationship behaviors without teaching them any specific skills (AWARENESS; $n = 33$). When using the MAT, both husbands and wives in all four groups demonstrated drops in quality over the 3 years following the interventions, and couples in the treatment groups failed to demonstrate significant differences from couples in the NoTx group. When using the NMQ to model change in negative relationship qualities over time, the analyses also failed to identify any differences between the couples receiving treatment and those in the NoTx group. However, when using the PMQ scores to model linear change in positive relationship qualities over time, couples in the NoTx group demonstrated significantly sharper declines in positives than did couples in all three active treatment groups (PREP, CARE, and AWARENESS). The results suggest that positive relationship evaluations can change over time independently of negative evaluations and that using only a global measure of quality like the MAT might have obscured meaningful treatment results.

Extending this work further, Rogge and Fincham (2010) developed optimized measures of positive and negative relationship quality using a combination of exploratory factor

analyses and IRT in a sample of more than 1,600 college students. The authors asked respondents to rate their relationships on separate sets of 20 positive (e.g., enjoyable, pleasant, alive) and 20 negative (e.g., bad, empty, lifeless) adjectives, giving similar instructions to those that Fincham and Linfield (1997) used (e.g., ''Considering only the positive qualities of your relationship and ignoring the negative ones, evaluate your relationship on the following qualities''). Factor analyses supported two dimensions of evaluation that were moderately correlated with one another. The IRT analyses were used to identify the items most effective for assessing positive qualities (PRQ) and the items most effective for assessing negative qualities (NRQ). Hierarchical regression analyses showed that the PRQ-4 and NRQ-4 offered unique information beyond a four-item measure of global relationship quality (CSI-4) for understanding self-reported positive interactions, negative interactions, satisfaction with sacrifice, vengefulness toward partner, hostile conflict behavior, and disagreement tolerance. Furthermore, the PRQ-4 and NRQ-4 displayed distinct patterns of validity within those regressions, with the NRQ-4 being more strongly related to things like vengefulness and hostile conflict behavior and the PRQ-4 being more strongly related to satisfaction with sacrifice and disagreement tolerance.

The results continue to suggest that positive and negative relationship qualities represent two distinct (albeit related) dimensions of relationship quality, each with its own unique information to contribute in attempts to understand relationship functioning and relationship behavior. Unfortunately, such distinctions have been obscured by the widespread assumption in the existing literature that positive and negative relationship evaluations are simply opposite points on a single dimension and can therefore be assessed with a single scale. As with the concerns raised by the noise in measurement of the MAT and DAS, only time will reveal how forcing negative and positive evaluations onto a single scale might have obscured many potentially interesting and informative results.

From a theoretical perspective, this bidimensional conceptualization has important implications. For example, Fincham and Linfield (1997) have already shown it can be used to identify two groups of spouses (those high on both dimensions vs. those low on both dimensions) who behave differently but are indistinguishable on unidimensional measures of marital quality (scoring in the midrange). It also has the potential to yield a richer picture of paths toward relationship distress. For instance, a decrease in positive evaluation that precedes an increase in negative evaluation may be quite different from one in which both processes occur in tandem or one in which the negative increases first and is followed by a decrease in the positive. Finally, it is possible that relationship processes have distinct impacts on positive versus negative evaluations of relationships. Indeed, links between relationship processes and positive or negative relationship evaluations might vary by the relative strength of those positive and negative evaluations. In fact, such theoretical distinctions might help explain seemingly conflictual results, such as findings that hostile conflict behavior can be associated with poorer relationship quality over time (e.g., Rogge & Bradbury, 1999) or with improved quality over time (e.g., Gottman & Krokoff, 1989) in assessments of relationship quality with a unidimensional construct. Only by expanding the theoretical conceptualization to a two-dimensional model of relationship evaluations would it be possible to test such possibilities. Regardless of the outcome, there are problems with self-report that need to be addressed, an issue to which we now turn.

The Limits of Self-Report: Enter Implicit Measures

The limitations of self-report have been extensively documented (e.g., Stone et al., 2000). Such limitations include impression management, motivated distortion, and the limits of self-awareness. The first has been widely recognized in marital research and has given rise to the development of a measure of social desirability that is specific to marriage, the Marital Conventionalization Scale (Edmonds, 1967). This scale contains items that describe the marriage in an impossibly positive light portraying the marriage as perfect and meeting the respondent's every need. It correlates strongly (in the .50–.73 range) with numerous measures of marital adjustment and satisfaction (Fowers & Applegate, 1996). Edmonds (1967) argued that the social desirability bias in responses to assessment of marital satisfaction was unconscious and unintended and therefore involved ''fooling oneself rather than

fooling others'' (p. 682). Although researchers have tried to control for this contaminant in the assessment of marital quality using Edmond's (1967) scale, increased concern about what it actually measures (Fowers & Applegate, 1996) renders such efforts moot and stresses the need for an alternative approach to this problem.

Motivated distortion and the limits of self-awareness in self-report have received relatively less attention in marital and family research. The failure to come to terms with nonconscious processes and instead assume that spouses have access to whatever we ask them about is an important limitation of the marital literature. This assumption has been thoroughly repudiated in the literature on social cognition and may account for the dramatic expansion of implicit measures in behavioral and social science in recent years (see Wittenbrink & Schwarz, 2007). Implicit measures aim to assess attitudes (or constructs) that respondents may not be willing to report directly or of which they may not even be aware. Such measures provide an index of the construct of interest without directly asking for verbal reports and are therefore likely to be free of social desirability biases.

The two major implicit measures used in research are use of priming methods and the Implicit Association Test (IAT) (Greenwald, McGhee, & Schwartz, 1998). Priming methods focus on automatic activation of evaluation associated with the primed stimulus. This produces a processing advantage for evaluatively congruent targets and a disadvantage for evaluatively incongruent targets, as the response suggested by the prime must be inhibited (for a review, see Fazio, 2001). By measuring response latency to targets following priming, we gain information about the evaluation of the primed object. In a similar vein, the IAT, which involves sorting words, assumes that ''if two concepts are highly associated, the IAT's sorting tasks will be easier when the two associated concepts share the same response than when they require different responses'' (Greenwald & Nosek, 2001, p. 85). We turn to consider such sorting tasks.

Word-Sorting Association Tasks Assessing Implicit Attitudes

The IAT and its derivative, the Go/No-Go Association Task (GNAT) (Nosek & Banaji, 2001) have been used to assess constructs thought to be heavily influenced by self-report biases, such as racial stereotypes, self-esteem, and psychopathology (see De Houwer, 2002; Greenwald & Farnham, 2000; Mitchell, Nosek, & Banaji, 2003). In the IAT, participants are presented with four types of stimuli, one at a time in random order, and are asked to categorize those stimuli into a left- or right-hand response. For example, respondents might be asked to respond with the left hand for ''good'' words and the right hand for ''bad'' words while simultaneously being asked to sort words from the two opposing target categories into right-hand and left-hand categories. By alternately pairing stimuli from each target category with ''good'' and ''bad'' response options across different blocks of trials and then looking for relative differences in speed of performance, it is possible to quantify implicit attitudes. For example, Greenwald et al. (1998) measured implicit prejudice against Black people by subtracting relative differences in response latencies from a stereotype-compatible condition (grouping Black names with ''bad'' words and White names with ''good'' words) from relative differences in response latencies from a stereotype-incompatible condition (Black with ''good'' and White with ''bad''). In this example, larger discrepancies in performance between the two conditions (e.g., notably faster performance on the stereotype-compatible trials and poorer performance on the stereotype-incompatible trials) would suggest stronger implicit stereotypes.

Although the IAT in its original form offers a powerful methodology for assessing implicit attitudes, it is somewhat limited in that it assesses those implicit attitudes as a contrast between two opposing target categories (e.g., Black vs. White, flowers vs. insects). This makes the implicit attitudes assessed specific to the contrasting categories used, and contradictory implicit attitudes can be obtained for the same target category by changing the contrasting category used in the task (e.g., Mitchell et al., 2003). Thus, different results could be expected when assessing implicit attitudes toward romantic partners if those attitudes are assessed using an IAT that contrasts partners with selves, partners with strangers, or partners with friends. In contrast, the GNAT does not require a contrasting category to assess implicit sentiment toward a target.

The GNAT is a highly similar word-sorting task in which respondents are presented with a mixture of stimuli, one at a time in random order,

and are asked to sort them by pressing a key when target words appear (a go response) and to refrain from pressing the key when distracter words appear (a no-go response). By alternately pairing good or bad words with words specific to the category of interest (e.g., romantic partner) in separate blocks of trials and measuring how quickly and accurately a subject responds, it is possible to assess the subject's implicit attitude toward that category. In a series of six studies examining the validity of the GNAT as an alternative to the more traditional IAT, Nosek and Banaji (2001) demonstrated that the GNAT was able to demonstrate comparable differences in performance to those found with the IAT when contrasting a category with generally positive implicit associations (fruit) from a category with generally negative implicit associations (bugs). The authors were also able to demonstrate that the GNAT could be used to identify such positive or negative implicit attitudes toward a single target category without the use of a contrasting category, offering a significant practical advantage over the IAT. Nosek and Banaji (2001) further demonstrated that the GNAT could be set to use either response speed (response latencies) as the primary index of performance or response accuracy (during a much faster and more difficult task) as the primary index of performance—achieving comparable results with both paradigms. Recently, the GNAT showed predictive validity in measuring fear of spiders (Teachman, 2007).

Assessments of Implicit Attitudes in the Relationship Literature

Although not absent from marital research, studies making use of word-sorting tasks to assess implicit attitudes are a rarity. Implicit measures were first introduced to this field in a study that sought to use response latency or reaction time to show that time taken to make evaluative judgments of the partner and the marriage moderated the relation between relationship quality and expected partner behavior (Fincham, Garnier, Gano-Phillips, & Osborne, 1995). In one task, respondents were asked to sort a set of 48 words including four partner stimuli (e.g., ''your partner,'' ''your spouse,'' partner's first name) into good and bad categories and response times for the partner stimuli were considered an implicit measure of relationship satisfaction accessibility. In a second task, response latencies were measured for 7 self-report Likert items assessing relationship quality embedded within a larger set of 22 items. The results indicated that compared to slow responders, fast responders on either task showed significantly higher correlations of self-reported marital quality and expected partner behavior in an upcoming interaction. This suggested that knowing a person's accessibility to their relationship quality at an implicit level might be associated with greater insight into their relationship behaviors and dynamics. Even though this innovative study drew significant attention at the time (e.g., Baucom, 1995; Beach, Etherton, & Whitaker, 1995), the tasks used confounded the implicit assessment of underlying sentiment (through response latencies) with the conscious (self-reported) assessment of relationship quality, thereby potentially obscuring the unique information that an implicit assessment of attitudes toward a romantic partner might provide. The findings were also somewhat limited by the fact that the entire assessment was limited to a small set (four or seven) of critical trials. This stands in contrast to the 140 critical trials typically afforded by IAT or GNAT paradigms.

Implicit assessments have also appeared in the literature examining the inclusion of others in the self. Aron, Aron, Tudor, and Nelson (1991) asked respondents to rate themselves and their partners on a list of 90 stimuli. They then measured reaction times as the respondents sorted those 90 stimuli into ''me'' or ''not me'' categories. They argued that individuals who had incorporated their romantic partner into their own self-image should have shown delays on the subset of stimuli for which they rated themselves as different from their partners (and correspondingly faster times on the stimuli for which they rated themselves as similar to their partners). Thus, the authors used performance on those two types of stimuli in the me and not-me as an implicit assessment of cognitive interdependence or closeness. The results supported their hypotheses, and the authors were able to demonstrate that higher levels of implicit closeness were correlated with higher levels of self-reported closeness. Extending those results, Aron and Fraley (1999) demonstrated that implicit closeness as assessed with the me and not-me task helped predict changes in self-reported closeness over 3 months as well as relationship breakup over that same interval. Unfortunately, the me and not-me task was also

limited by the small number of stimuli that fell into the two critical categories for individual respondents—with some respondents having no stimuli or one stimulus falling into the self-different-from-partner categories.

More recently, the IAT was used to measure implicit attitudes toward romantic partners in two published studies. Zayas and Shoda (2005) used the name that respondents used to refer to their romantic partner as the category label (e.g., ''John'') and the contrasting category label (e.g., ''Not-John'') for the two target categories to be alternately paired with ''good'' and ''bad'' responses. They then used a set of unique descriptive words (e.g., nickname, hair color, city of birth) as partner stimuli and a set of words unassociated with each respondent's partner as not-partner stimuli to be mixed in with the ''good'' and ''bad'' stimuli. Their results demonstrated that, at a cross-sectional level, more positive implicit attitudes as assessed by the partner-IAT were associated with higher levels of secure attachment (a healthy internal working model of romantic relationships) and lower levels of attachment avoidance (discomfort with emotional closeness and intimacy). Banse and Kowalick (2007) used a similar type of partner IAT to examine its association with concurrent relationship quality and well-being. They recruited women who were living in a shelter, women who were hospitalized because of pregnancy complications, women who had recently fallen in love, and female students. Their results showed that abused women demonstrated more negative implicit attitudes toward partners and that positive implicit attitudes were associated with higher levels of secure attachment across all women in the sample. Unexpectedly, implicit attitudes were not significantly associated with self-reported relationship quality.

Extending this work, Lee, Rogge, and Reis (in press) sought to develop an implicit measure of attitudes toward romantic partners independent of explicit (self-report) evaluations and without the need for a contrasting category to that of ''partner.'' The authors also sought to validate this implicit measure over a much longer time frame—during which greater amounts of relationship change could be expected to occur. Finally, extending the groundbreaking work with self-report measures of relationship quality, the authors sought to disentangle positive and negative implicit attitudes toward a romantic partner, examining them as separate dimensions with potentially separate predictive validities. Thus, Lee, Rogge, and Reis (in press) examined the unique validity of a partner-focused GNAT to predict relationship dissolution over 12 months across two studies. They asked respondents to provide three stimuli specific to their romantic partners (e.g., name, nickname, distinguishing characteristic). The respondents then completed a GNAT in which the partner stimuli were alternately paired with ''good'' and ''bad'' words as the target stimuli across two separate blocks of 70 trials. Given some of the inherent limitations of using reaction time data (e.g., extreme skew, high noise-to-signal ratios), the task was set up as a rapid task using accuracy as the primary measure of performance. Higher levels of performance on the trials in which partner stimuli were paired with good stimuli as targets (partner-good trials) were hypothesized to reflect stronger positive implicit attitudes toward a romantic partner, and higher levels of performance on the partner-bad trials were hypothesized to reflect stronger negative implicit attitudes toward a partner. After completing the partner-GNAT, respondents completed a battery of self-report questionnaires assessing relationship quality, hostile conflict behavior, and neuroticism. The respondents were then contacted four to six times over the following 12 months to assess relationship stability.

Interestingly, in both samples, performance on the partner-good and partner-bad trials demonstrated positive correlations with each other ($r = .46$), despite being hypothesized to reflect positive and negative implicit attitudes, respectively. This likely represents shared method variance due to the common mechanics of the word-sorting task (e.g., general levels of ability, effort expended on the task, ability to sustain attention, comfort with using computers). To control for this, the partner-good and partner-bad performance indices were entered pairwise in all subsequent analyses to ensure that their shared variance would be dropped from the models. Discrete-time hazard modeling in a hierarchical linear modeling (HLM) (Raudenbush & Byrk, 2002) framework demonstrated that lower levels of performance on partner-good trials was associated with significantly higher risk for breakup over the following 12 months in both samples, even after controlling for self-reported relationship quality, hostile conflict and neuroticism. There was also partial support to

suggest that performance on the partner-bad trials had unique predictive validity. In one of the samples, performance on the partner-bad trials demonstrated a significant main effect such that a higher level of performance on partner-bad trials was associated with higher risk of breakup. In the other sample, performance on partner-good and partner-bad trials interacted such that it was specifically the individuals with below average partner-good performance and above average partner-bad performance who were at greatest risk of breaking up over the following 12 months. Taken as a set, the findings of Lee, Rogge, and Reis (in press) suggest that implicit assessments of positive and negative attitudes toward a romantic partner do offer insight into the functioning of those romantic relationships that cannot be obtained through traditional self-report scales. The results also lend additional support to conceptualizing relationship quality as two distinct dimensions, which suggests that individuals can have relatively distinct positive and negative implicit evaluations of romantic partners—each with unique information to contribute in understanding relationship stability over time.

Although the implicit measures described are promising, marital and family scholars have not gravitated toward them. This could be because they require specific equipment and use paradigms that are not commonly found in the family literature. It is therefore worth noting that paper and pencil (low-tech) implicit measures have been developed but have been used sparingly (see Vargas, Sekaquaptewa, & Hippel, 2007) relative to their high-tech counterparts. We now report on a low-tech approach to assessing relationship quality implicitly.

New Wine, Old Wineskin

As Fazio and Olson (2003, p. 303) noted, modern implicit measures that assess constructs without directly asking about them are, in this regard, no different from ''earlier proposals regarding projective methods'' (see Proshansky, 1943). This is not the context to reiterate the pros and cons of projective techniques. Instead, we describe a stunningly simple way in which we try and get at relationship quality using what cannot be considered anything but a projective method. Specifically, we have been asking study participants to do the following: ''Please draw a picture with (a) a house, (b) a tree (c) a car and (d) two people *who represent you and your partner*. You may draw them in any way you like, but you must include the above items. Please label the figure that represents you as 'me' and the one that represents your partner as 'partner.''' We then used the distance between the necks of each person in the drawing (measured in millimeters) as an index of relationship quality.

The results obtained have been quite extraordinary. Across two samples, there was good evidence of convergent validity. For example, neck distance correlated significantly with Funk and Rogge's (2007) CSI scores, with the likability of the partner and commitment to him or her. Importantly, the neck-distance measure predicted a number of relevant variables 4 weeks later over and beyond the CSI and initial level of the variable predicted. The variables thus predicted included expression of appreciation, commitment to partner, likability of partner, intimacy, mattering, perceived relationship maintenance efforts of the partner, and perceived commitment of partner. Finally, the neck measure may also foretell extradyadic sexual behavior and how safe a person feels in the relationship ($p < .06$ in each case).

The results are preliminary but deserve mention because they remind us that there is potentially much to gain from exploring some very old and well-known methods as we seek implicit measures of relationship quality.

Coda

Although implicit measures have been widely used, enthusiasm for them has not been matched by theoretical development. It is therefore important to note that we are not advocating use of such measures for their own sake. Rather, we strongly agree with Fazio and Olson (2003) that, ''when their application, use, and interpretation is guided by relevant theory and past literature, implicit measures have the potential to serve as useful methodological tools for testing hypotheses'' (p. 320).

Conclusion

We have traversed a great deal of territory in this article. Rather than attempt to offer an exhaustive analysis of each topic addressed, our goal has been to pique the reader's interest and provide some pivotal citations for further reading on the topic.

It appears that the marital literature is at a crossroads in regard to its most frequently studied construct, relationship quality. The weight of inertia has promulgated use of measures that lack conceptual clarity and can even be questioned on psychometric grounds. This continuing practice has stunted theory development. After documenting that case, we offered a conceptually simple and theoretically advantageous view of relationship quality as evaluation of the relationship. We then illustrated how this conceptualization can be pursued in the context of modern test theory providing, en passant, a brief introduction to IRT. Next, we offered an expansion of the unidimensional view of relationship quality, suggesting that positive and negative evaluations might be conceptually distinct and should be assessed separately. We then presented psychometric data supporting such a theoretical view and outlined some of the most important theoretical implications, including the ability to identify different groups of spouses who tend to fall near the midpoint of unidimensional relationship quality scales but behave quite differently from each other. Recognizing the limitations of self-reported relationship quality, we drew on related attitude research and developments to derive implicit attitude measures. Again, after offering a brief introduction to such measures, we offered examples of their application to the assessment of relationship quality, further supporting a more complex two-dimensional conceptualization of relationship quality. Thus, in this article, we strove to demonstrate how theory could shape psychometric inquiries and how psychometrics could, in turn, help refine theoretical development.

Clearly, there is a choice to be made. Either we allow inertia to condemn us to a future that is much like the past, or we break out of our comfort zone and pursue new approaches to conceptualizing and measuring relationship quality in the dominant epistemology of research on this topic. As evidenced by the other articles in this special issue, the construct of relationship quality has come under attack on fundamental philosophical grounds. We would argue that this construct is still fundamentally sound, as people seem to inherently experience relationships on globally positive and negative dimensions—both at a conscious level and at an implicit level, of which they might not be fully aware. However, if this construct is to withstand the theoretical and philosophical criticisms levied against it (and to remain a useful tool for advancing our knowledge of relationships), then it is necessary for couples researchers to adopt new methods of conceptualizing and measuring relationship quality. It is not easy or comfortable to explore literatures in other disciplines, especially when it involves mastering new and sometimes complex methods. However, failure to do so brings with it a high cost and one that will potentially lead our field to collapse under the weight of its own conceptual confusion. Such a future is simply too ghastly to contemplate.

Note

This article was made possible by grant 90FE0022/01 from the Department of Health and Human Services Administration for Children and Families awarded to the first author. The authors thank Sesen Negash and Natalie Sentore for their comments on an earlier draft of this article.

References

Aron, A., Aron, E. N., Tudor, M., & Nelson, G. (1991). Close relationships as including other in the self. *Journal of Personality and Social Psychology*, *60*, 241 – 253.

Aron, A., & Fraley, B. (1999). Relationship closeness as including other in the self: Cognitive underpinnings and measures. *Social Cognition*, *17*, 140 – 160.

Banse, R., & Kowalick, C. (2007). Implicit attitudes towards romantic partners predict well-being in stressful life conditions: Evidence from the antenatal maternity ward. *International Journal of Psychology*, *42*, 149 – 157.

Baucom, D. H. (1995). A new look at sentiment override—Let's not get carried away yet: Comment on Fincham et al. (1995). *Journal of Family Psychology*, *9*, 15 – 18.

Beach, S. R., Etherton, J., & Whitaker, D. (1995). Cognitive accessibility and sentiment override— Starting a revolution: Comment on Fincham et al. (1995). *Journal of Family Psychology*, *9*, 19 – 23.

Bradbury, T. N., & Fincham, F. D. (1987). The assessment of affect in marriage. In K. D. O'Leary (Ed.), *Assessment of marital discord: An integration for research and clinical practice* (pp. 59 – 108). Hillsdale, NJ: Erlbaum.

Dahlstrom, W. G. (1969). Recurrent issues in the development of the MMPI. InJ. M. Butcher (Ed.), *Research developments and clinical applications* (pp. 1 – 40). New York: McGraw Hill.

De Houwer, J. (2002). The Implicit Association Test as a tool for studying dysfunctional associations in psychopathology: Strengths and limitations. *Journal of Behavior Therapy and Experimental Psychiatry*, *33*, 115 – 133.

Edmonds, V. H. (1967). Marital conventionalization: Definition and measurement. *Journal of Marriage and the Family*, *24*, 349 – 354.

Fazio, R. H. (2001). On the automatic activation of associated evaluations: An overview. *Cognition and Emotion*, *15*, 115 – 141.

Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and use. *Annual Review of Psychology*, *54*, 297 – 232.

Fincham, F. D. (1998). Child development and marital relations. *Child Development*, *69*, 543 – 574.

Fincham, F. D., & Beach, S. R. (1999). Marital conflict: Implications for working with couples. *Annual Review of Psychology*, *50*, 47 – 77.

Fincham, F. D., & Bradbury, T. N. (1987). The assessment of marital quality: A reevaluation. *Journal of Marriage and the Family*, *49*, 797 – 809.

Fincham, F. D., Garnier, P. C., Gano-Phillips, S., & Osborne, L. N. (1995). Preinteraction expectations, marital satisfaction, and accessibility: A new look at sentiment override. *Journal of Family Psychology*, *9*, 3 – 14.

Fincham, F. D., & Linfield, K. J. (1997). A new look at marital quality: Can spouses feel positive and negative about their marriage? *Journal of Family Psychology*, *11*, 489 – 502.

Fowers, B. J., & Applegate, B. (1996). Marital satisfaction and conventionalization examined dyadically. *Current Psychology*, *15*, 197 – 214.

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

Glenn, N. D. (1990). Quantitative research on marital quality in the 1980s: A critical review. *Journal of Marriage and the Family*, *52*, 818 – 831.

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

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

Gottman, J. M., & Levenson, R. W. (1984). Why marriages fail: Affective and physiological patterns in marital interaction. In J. C. Masters & K. Yarkin-Levin (Eds.), *Boundary areas in social and developmental psychology* (pp. 67 – 106). New York: Academic Press.

Greenwald, A. G., & Farnham, S. D. (2000). Using the Implicit Association Test to measure self-esteem and self-concept. *Journal of Personality and Social Psychology*, *79*, 1022 – 1038.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, *74*, 1464 – 1480.

Greenwald, A. G., & Nosek, B. A. (2001). Health of the Implicit Association Test at age 3. *Zeitschrift für Experimentelle Psychologie*, *48*, 85 – 93.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Hendrick, S. S. (1988). A generic measure of relationship satisfaction. *Journal of Marriage and the Family*, *50*, 93 – 98.

Heyman, R .E., Sayers, S. L., & Bellack, A. S. (1994). Global marital satisfaction versus marital adjustment: An empirical comparison of three measures. *Journal of Family Psychology*, *8*, 432 – 446.

Jacobson, N. S. (1985). The role of observation measures in marital therapy outcome research. *Behavioral Assessment*, *7*, 287 – 308.

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

Karney, B. R., & Bradbury, T. N. (1997). Neuroticism, marital interaction, and the trajectory of marital satisfaction. *Journal of Personality and Social Psychology*, *72*, 1075 – 1092.

Lee, S., Rogge, R. D., & Reis, H. T. (In press). Assessing the seeds of relationship decay: Using implicit evaluations to detect the early stages of disillusionment. *Psychological Science.*

Locke, H. J., & Wallace, K. M. (1959). Short marital adjustment prediction tests: Their reliability and validity. *Marriage and Family Living*, *21*, 251 – 255.

Mattson, R. E., Paldino, D., & Johnson, M. D. (2007). The increased construct validity and clinical utility of assessing relationship quality using separate positive and negative dimensions. *Psychological Assessment*, *19*, 146 – 151.

Mitchell, J. P., Nosek, B. A., & Banaji, M. R. (2003). Contextual variations in implicit evaluation. *Journal of Experimental Psychology: General*, *132*, 455 – 469.

Norton, R. (1983). Measuring marital quality: A critical look at the dependent variable. *Journal of Marriage and the Family*, *45*, 141 – 151.

Nosek, B. A., & Banaji, M. R. (2001). The go/no-go association task. *Social Cognition*, *19*, 625 – 664.

Proshansky, H. M. (1943). A projective method for the study of attitudes. *Journal of Applied Social Psychology*, *38*, 393 – 395.

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

Rogge, R. D., & Bradbury, T.N. (1999). Till violence does us part: The differing roles of communication and aggression in predicting adverse martial outcomes. *Journal of Consulting and Clinical Psychology*, *67*, 340 – 351.

Rogge, R. D., Cobb, R. J., Johnson, M. D., Lawrence, E. E., & Bradbury, T. N. (2002). The CARE

program: A preventive approach to marital intervention. In A. S. Gurman & N. S. Jacobson (Eds.), *Clinical handbook of couple therapy* (3rd ed., pp. 420 – 435). New York: Guilford Press.

Rogge, R .D., Cobb, R. J., Lawrence, E. E., Johnson, M. D., Story, L. B., Rothman, A. D., et al. (2010). *Teaching skills vs. raising awareness: The effects of the PREP, CARE and AWARENESS programs on 3-year trajectories of marital functioning*. Unpublished manuscript, University of Rochester, Rochester, NY.

Rogge, R. D., & Fincham, F. D. (2010). *Disentangling positive and negative feelings toward relationships: Development and validation of the Positive-Negative Relationship Quality (PNRQ) scale*. Unpublished manuscript, University of Rochester, Rochester, NY.

Samejima, F. (1997). Graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85 – 100). New York: Springer.

Schumm, W. A., Nichols, C. W., Schectman, K. L., & Grinsby, C. C. (1983). Characteristics of responses to the Kansas Marital Satisfaction Scale by a sample of 84 married mothers. *Psychological Reports*, *53*, 567 – 572.

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

Stone, A. A., Turkan, J. S., Bachrach, C. A., Jobe, J. B., Kurtzman, H. S., & Cain, V. S. (2000). *The science of self-report: Implications for research and practice.* Mahwah, NJ: Erlbaum.

Teachman, B. A. (2007). Evaluating implicit spider fear associations using the go/no-go association task. *Journal of Behavior Therapy and Experimental Psychiatry*, *38*, 157 – 167.

Trost, J. *E.* (1985). Abandon adjustment! *Journal of Marriage and the Family*, *47*, 1072 – 1073.

Vargas, P.T., Sekaquaptewa, D., & Hippel, W. (2007). Armed only with paper and pencil: ''Low-tech'' measures of implicit attitudes. In B. Wittenbrink & N. Schwarz (Eds.), *Implicit measures of attitudes* (pp. 103 – 124). New York: Guilford Press.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scale. *Journal of Personality and Social Psychology*, *54*, 1063 – 1070.

Watson, D., Clark, L. A., Weber, K., Assenheimer, J. S., Strauss, M. E., & McCormick, R. A. (1995). Testing a tripartite model: II. Exploring the symptom structure of anxiety and depression in student, adult, and patient samples. *Journal of Abnormal Psychology*, *104*, 15 – 25.

Weiss, R. L. (1980). Strategic behavioral marital therapy: Toward a model for assessment and intervention. In J. P. Vincent (Ed.), *Advances in family intervention, assessment and theory* (Vol. 1, pp. 229 – 271). Greenwich, CT: JAI Press.

Wittenbrink, B., & Schwarz, N. (2007). *Implicit measures of attitudes*. New York: Guilford Press.

Zayas, V., & Shoda, Y. (2005). Do automatic reactions elicited by thoughts of romantic partner, mother, and self relate to adult romantic attachment? *Personality and Social Psychology Bulletin*, *31*, 1011 – 1025.

Appendix

For a simple dichotomous question (true or false), the two-parameter logistic model would take the form: $P_i(\theta) = 1/[1 + exp(-\alpha_i\ \theta)\ (\theta - \beta_i)]$. In this model, $P_i(\theta)$ indicates the probability that an individual with trait level θ will endorse item i as true. Thus, IRT would estimate two-item parameters for the dichotomous item (α_i and β_i)—the discrimination coefficient (α) estimates the relative amount of information an item contributes and the difficulty coefficient (β) estimates the region of the latent trait where the item is most informative. The characteristics of a response curve can then be synthesized with the following equation to estimate the overall amount of information the item provides: $I_i(\theta) = [P_i'(\theta)]^2/[P_i(\theta)(1 - P_i[\theta])]$, where $P_i'(\theta)$ is the first derivative of $P_i(\theta)$ with respect to θ. The $I_i(\theta)$ function creates an item information curve (IIC), revealing the relative amount of information the item contributes at various points along the continuum of the underlying trait (θ). As mentioned earlier, the standard error of an item is simply $SE(\theta) = 1/s\theta rt[I\theta(\theta)]$, or basically the inverse of the information provided, so these estimates of information also serve as estimates of the precision of measurement (the lack of error in measurement). Information curves of individual items can be summed to create test information curves (TICs), which estimate the information (and precision) a set of items provides to the assessment of the underlying trait. Because IRT information curves are based on precise estimates of each respondent's latent score, with a sufficiently large and diverse sample, the estimated information curves become sample independent—accurately estimating how sets of items will perform in any number of new samples. An IRT analysis of the existing measures of relationship quality would therefore allow researchers to create TICs for those measures, placing them on the same ruler to

determine which scales measure up to the task of assessing relationship quality.

To apply IRT analyses to items with Likert response scales, one can use the Graded Response Model (GRM) (Samejima, 1997). In the GRM, an item with m response options is considered a set of m—1 dichotomous thresholds. Thus, for an item with four response choices, GRM would model response curves (operating characteristic curves, or OCCs) for the three dichotomous decisions respondents face (1 vs. 2, 2 vs. 3, and 3 vs. 4). The three OCCs are assumed to have the same discrimination (α_i) parameter but three distinct difficulty (β_{ij}) parameters. Therefore, GRM affords the possibility of very precisely examining the quality of information provided by items with Likert response scales across a range of possible relationship quality values (typically from −3 to 3 SD around the population mean).