Psychological Assessment
2017, Vol. 29, No. 8, 1028–1043

1040-3590/17/$12.00 http://dx.doi.org/10.1037/pas0000392

Positive and Negative Evaluation of Relationships: Development and Validation of the Positive–Negative Relationship Quality (PN-RQ) Scale

Ronald D. Rogge
University of Rochester

Frank D. Fincham
Florida State University

Dev Crasta
University of Rochester

Michael R. Maniaci
Florida Atlantic University

Three studies were undertaken to develop the Positive–Negative Relationship Quality scale (PN-RQ), conceptualizing relationship quality as a bidimensional construct in which the positive qualities of a relationship are treated as distinct from its negative qualities. Analyses in emerging adults (Study 1: N = 1,814), in online respondents (Study 2: N = 787) with a 2-week follow-up, and in a single group pre-intervention–post-intervention study (Study 3: N = 54) of the Promoting Awareness, Improving Relationships (PAIR) program provided support for (a) positive and negative qualities as distinct dimensions via confirmatory factor analysis (CFA), (b) the PN-RQ representing an item response theory-optimized measure of these 2 dimensions, (c) substantive differences between indifferent (low positive and negative qualities) and ambivalent (high positive and negative qualities) relationships potentially obscured by unidimensional scales, (d) high levels of responsiveness of the PN-RQ scales to change over time, (e) the unique predictive validity offered over time by the PN-RQ scores beyond that offered by scores of current unidimensional measures of relationship quality, and (f) the unique longitudinal information gained by using the PN-RQ as a bidimensional outcome measure in an intervention study. Taken together, the studies offer promising support for the PN-RQ scales suggesting that they have the potential to advance both basic and applied research.

Keywords: couples, marriage, satisfaction, item response theory, measure development

Relationship quality has served as a central construct in romantic relationship research (see Fincham, Rogge, & Beach, in press; Karney & Bradbury, 1995, for reviews) possibly because it is viewed as the final common pathway that leads to relationship breakdown (Jacobson, 1985). As a result, researchers have attempted to understand not only the factors that keep couples together but also the factors that keep their relationships fulfilling and rewarding. Traditionally, this construct has been conceptualized as a single dimension (i.e., extremely dissatisfied to extremely satisfied) and has therefore been measured with self-report scales that are summed into a single total scores (e.g., the Marital Adjustment Test [MAT]; Locke & Wallace, 1959], the Dyadic Adjustment Scale [DAS]; Spanier, 1976; the Couples Satisfaction Index [CSI]; Funk & Rogge, 2007). Recent work has challenged such a unidimensional conceptualization (e.g., Fincham & Rogge, 2010), suggesting that assessing positive and negative evaluations of a relationship separately can yield greater insights into relationship functioning (e.g., Fincham & Linfield, 1997; Mattson, Rogge, Johnson, Davidson, & Fincham, 2013) and might even uncover treatment effects obscured by unidimensional scales (see Fincham & Bradbury, 1987; Rogge, Cobb, Lawrence, Johnson, & Bradbury, 2013). The current study built on this line of work by using Item Response Theory (IRT; Hambleton, Swaminathan, & Rogers, 1991) analyses on larger item pools to develop optimized bidimensional scales to assess positive and negative relationship quality.

Conceptual Definition of Relationship Quality

Despite its centrality close relationship research, the construct of relationship quality has been somewhat mired by a lack of conceptual clarity as evidenced by the diverse range of terms that have been used interchangeably to label it: satisfaction, adjustment, success, happiness, companionship, closeness, bond, as well as other synonyms of the term quality (see Fincham & Rogge, 2010, for a review). These terms loosely correspond to differing conceptual approaches to operationalize this construct ranging from a broad focus on a range of interpersonal processes assessing relationship *adjustment* (e.g., affection, companionship, conflict; yielding heterogeneous scales) to a narrow focus on global evaluations of relationship *quality* (e.g., satisfaction, happiness, bond).

This article was published Online First October 13, 2016.

Ronald D. Rogge, Department of Clinical and Social Sciences in Psychology, University of Rochester; Frank D. Fincham, The Family Institute, Florida State University; Dev Crasta, Department of Clinical & Social Sciences in Psychology, University of Rochester; Michael R. Maniaci, Department of Psychology, Florida Atlantic University.

The studies were funded by internal grants at the University of Rochester and Florida State University. We thank the participants who completed our studies. This research would not have been possible without their support.

Correspondence concerning this article should be addressed to Ronald D. Rogge, Department of Clinical and Social Sciences in Psychology, University of Rochester, 462 Meliora Hall–RC Box 270266, Rochester, NY 14627-0266. E-mail: rogge@psych.rochester.edu

Although the scores of scales created from these differing approaches show strong correlations with one another and a high level of convergent validity (e.g., Funk & Rogge, 2007), recent measurement work using IRT (a large-sample technique for critically evaluating precision of measurement) has favored a narrower conceptual focus as it tends to yield scales with greater precision (Funk & Rogge, 2007). Thus, the current study focused conceptually on individuals' global subjective evaluations of their relationships and used the term *relationship quality* to label that construct. Given its common conceptual focus and its widespread prevalence in the couples and marital literatures (e.g., Karney & Bradbury, 1995; Funk & Rogge, 2007), we acknowledge that the term *relationship satisfaction* is synonymous with *relationship quality* and can be used interchangeably. Having said that, we prefer to use the term *relationship quality* as it more clearly encompasses both positive and negative evaluations.

Advantages of a Two-Dimensional Conceptualization of Relationship Quality

The current research is informed by the view that individuals in romantic relationships simultaneously hold both negative and positive sentiments toward romantic partners (somewhat independently). Consequently, constraining the assessment of relationship quality to a single dimension could be obscuring important phenomena and oversimplifying theories (see Fincham & Rogge, 2010). The current study sought to create a measure assessing positive and negative evaluations of relationships as distinct yet related constructs that jointly represent *relationship quality*. This perspective is consistent with robust findings in the affect literature, exemplified by scales like the Positive and Negative Affect Scale (PANAS; Watson, Clark, & Tellegen, 1988) and the Mood and Symptom Questionnaire (MASQ; Watson et al., 1995), suggesting that the experience of positive and negative affect (or distress & vitality) are substantively distinct yet related phenomena, best assessed separately. This two-dimensional conceptualization is also consistent with larger body of research that positive and negative components of an array of processes (e.g., motivation, affect, cognitive evaluation, personality) might form more general appetitive and aversive behavioral systems that are meaningfully distinct yet related to each other (Gable, Reis, & Elliot, 2003). Fincham and Linfield (1997) first explored the potential benefits of a bidimensional conceptualization of relationship quality by developing the Positive and Negative Qualities in Marriage Scale (PN-QIMS). The PN-QIMS consists of two 3-item subscales assessing positive and negative qualities separately. 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*"). Analyses in a sample of 123 married couples demonstrated that the PN-QIMS subscales each accounted for unique predictive variance in self-reports of conflict behavior and attributions even after controlling for MAT scores. Furthermore, differences emerged between indifferent individuals (those perceiving low negative qualities and low positive qualities) and ambivalent individuals (those perceiving high negative qualities and high positive qualities) – two groups of participants that could not be distinguished by a unidimensional measure of relationship quality (e.g., the MAT). Extending this work, Mattson, Paldino, and Johnson (2007) demonstrated that the PN-QIMS accounted for unique predictive variance in objectively coded positive and negative behavior during problem discussions, even after controlling for a unidimensional measure of quality.

To enhance the assessment of positive and negative relationship qualities, Mattson and colleagues (2013) drew upon the work of Osgood (1964) that explored the cross-cultural meanings of opposing adjective pairs (e.g., *good* vs. *bad*, *strong* vs. *weak*), developing the semantic differential as a measure of global evaluations. Specifically, the semantic differential asks subjects to rate a target (e.g., a relationship) on a Likert scale between opposing pairs of adjectives. Exploratory factor analyses within six different cultural contexts (i.e., American, Dutch, Finnish, Flemish, Japanese, Canadian; Osgood, 1964) identified three common dimensions of adjective pairs across cultures: evaluation pairs (e.g., *good–bad*, *pleasant–unpleasant*, *enjoyable–miserable*), activity pairs (e.g., *alive–lifeless*, *active–passive*, *invigorating–draining*), and potency pairs (e.g., *strong–weak, full–empty, deep–shallow*). Thus, the conceptual construct of global evaluations, when assessed with adjective pairs, could be further subdivided into three, closely related types of items assessing a common construct. A 15-item semantic differential assessing *relationship quality* with items spanning these three domains was shown to be a highly precise unidimensional measure, offering more information than widely used, heterogeneous measures of *relationship adjustment* (e.g., the MAT and the DAS; Funk & Rogge, 2007). To build on this work, Mattson and colleagues (2013) took the 7 semantic differential items that had been identified with IRT to be highly effective and precise measures of relationship quality (Funk & Rogge, 2007) and split the adjective pairs into separate positive and negative scales creating the Positive–Negative Semantic Differential (PN-SMD): a 14-item measure that asks participants to rate the qualities of their relationships on seven positive adjectives (e.g., good, enjoyable) and separately on seven negative adjectives (e.g., bad, miserable). Results across two samples showed that the PN-SMD offered unique predictive information beyond that offered by unidimensional measures of relationship quality—providing more nuanced insights into current relationship quality and into trajectories of relationship quality over 18 months by shifting to a bidimensional conceptualization of quality. Extending these results to the evaluation of treatment effects over time, Rogge and colleagues (2013) examined change in relationship quality over 3 years in a sample of 174 newlywed couples who received one of 4 treatment conditions. Modeling change in positive and negative relationship qualities with the PN-QIMS revealed differences among the treatment conditions that failed to emerge with the MAT. Taken as a set, this growing body of work suggests that unique predictive variance can be gained by disentangling the assessment of positive and negative relationship qualities with the use of a bidimensional conceptualization of relationship quality.

Using Item Response Theory to Optimize the Assessment of Relationship Quality

Classic approaches to the development of self-report scales have primarily relied upon correlational techniques like exploratory factor analysis (EFA), CFA, and Cronbach's alpha coefficients (see Clark & Watson, 1995, for an overview). These techniques can be highly effective, particularly within fairly small samples (e.g., 100–300 subjects), at creating internally consistent scales. More recently, re-

searchers have augmented these approaches with the use of IRT (Hambleton et al., 1991) to develop psychometrically optimized scales by maximizing precision and minimizing measurement noise (e.g., Fraley, Waller, & Brennan, 2000; Funk & Rogge, 2007; Shaw & Rogge, 2016). IRT accomplishes this by estimating latent scores (termed θ in IRT equations) for each subject on the construct being examined in an analysis in much the same way that Structural Equation Modeling (SEM) estimates latent scores. IRT then examines the response curves of each item to determine if subjects with higher θ scores select higher response options and subjects with lower θ scores select lower response options. To the degree that this is true for a specific item, it is considered to be an effective and informative item for assessing θ. Thus, IRT provides estimates of the discriminating information that specific items can provide to a scale. Although IRT requires much larger sample sizes, when conducted in appropriately large and diverse samples, the results of IRT become sample-invariant, helping to reveal how items and scales will operate in a wide range of future samples.

IRT has been used to augment traditional measurement analyses (e.g., EFA, CFA) to create psychometrically optimized self-report scales, including measures of adult attachment (the Experiences in Close Relationships—Revised scales of Fraley et al., 2000), global relationship satisfaction (the Couples Satisfaction Index or CSI scales of Funk & Rogge, 2007), and sexual quality (the Quality of Sex Inventory or QSI scales of Shaw & Rogge, 2016). This approach typically involves starting with a large item pool, using correlational analyses like EFA to identify sets of unidimensional items (assessing a single construct) and then using IRT to identify the most informative items. Results with these IRT-optimized scales suggest they offer greater precision and power for detecting differences (e.g., Funk & Rogge, 2007).

The Current Study

We sought to advance the work on bivariate measures of relationship quality by developing an IRT-optimized measure, the Positive–Negative Relationship Quality (PN-RQ) Scale.

Study 1: To diversify the assessment of positive and negative relationship quality, a large sample (N = 1,814) of emerging adults rated the quality of their relationships on a set of 20 positive and 20 negative adjectives spanning the three primary dimensions of positive and negative adjectives identified by Osgood's (1964) groundbreaking work: evaluation, activity, and potency. We then used CFA and IRT analyses to develop PN-RQ scale.

Study 2: To validate the PN-RQ against existing bivariate measures and explore its longitudinal properties, we had a large (N = 787) sample of online respondents complete the PN-RQ alongside existing bidimensional measures (the PN-SMD, the PN-QIMS) and unidimensional measures (the CSI) of relationship quality at two waves. This allowed us to: directly compare the discriminating information provided by the various relationship quality scales (using additional IRT analyses), evaluate the responsiveness to change of the scales (using Minimal Detectible Change or MDC_{95} estimates; see Stratford et al., 1996), and quantify the unique predictive validity of PN-RQ scores across 2 weeks.

Study 3: To explore the potential utility of the PN-RQ in the context of a single-group treatment study, we collected baseline and 1-month PN-RQ and CSI scores from 74 individuals engaging the PAIR program, a self-guided intervention encouraging couples to use specific movies (e.g., *American Beauty*) with a set of semistructured discussion questions as a nonthreatening way to engage in discussions of their own relationships. An earlier version of the PAIR intervention was associated with lower separation/divorce rates over the first three years of marriage (Rogge et al., 2013).

Study 1

Method

Participants. Participants were 1,814 undergraduate students completing an introductory family relations course that contained students representing all colleges and majors at a South Eastern university (Fincham, Cui, Braithwaite, & Pasley, 2008), representing 96% of the students in the course invited to participate. The participants were predominantly female (77%) and Caucasian (72%) with 14% African American, 11% Latino and 3% Asian. A majority of the respondents (54%) were in romantic relationships (76% in exclusive dating relationships, 21% in nonexclusive dating relationships, 2% engaged and 1% married) and completed the relationship measures with respect to their romantic partners. The remaining respondents completed the relationship measures with respect to a close relationship: 38% reporting on friends, 5% on family members, and 3% on roommates. For the respondents in romantic relationships, the average length of relationships was 1.5 years (SD = 1.3).

Procedure. The methods for all studies reported were approved by university institutional review boards (IRBs). The students participating in Study 1 were offered multiple options to earn class credit—one being the opportunity to participate in this study via an online survey.

Measures

Global relationship satisfaction. Respondents completed the 4-item Couples Satisfaction Index (CSI-4; Funk & Rogge, 2007). The items were rated on 6- and 7-point Likert scales, and were summed so that higher scores reflected higher levels of global satisfaction. Responses to the items demonstrated high internal consistency both with a traditional Cronbach's coefficient (α = .92) and with coefficient ω (ω = .930, 95% confidence interval [CI] [.921, .937], conducted per Dunn, Baguley, & Brunsden, 2014) an index with advantages over α given its focus on how much the items of a scale measure *one common factor* (see Revelle & Zinbarg, 2009).

Negative interaction. Negative interaction was measured by 4 items from The Communication Warning Signs Scale (Stanley & Markman, 1997; e.g., *little arguments escalate into ugly fights with accusations, criticisms, name calling, or bringing up past hurts*). Responses ranged from *never or almost never* (0) to *frequently* (2), and were summed so higher scores reflected more negative interaction (α = .73; ω = .736, 95% CI [.710, .760]).

Positive interaction. Positive interaction was measured by 2 positive items from The Communication Warning Signs Scale (Stanley & Markman, 1997): "*we have a lot of fun together*," "*we have great conversations where we just talk as good friends*."

Responses ranged from *strongly disagree* (0) to *strongly agree* (4) and were summed so that higher scores reflected more positive interaction ($\alpha = .78$; $\omega = .782$, 95% CI [.738, .817]).

Unforgivingness. The nine item Relationship Forgiveness Scale (Fincham, Beach, & Davila, 2004) was used to assess individual's forgiveness following a transgression in a close relationship (e.g., *when my partner wrongs or hurts me I: find a way to make her/him regret it*). Responses ranged from *strongly disagree* to *strongly agree* (0 to 5) and were summed so that higher scores indicated higher levels of unforgivingness—thereby retaining the original valence of the items of the scale ($\alpha = .83$; $\omega = .825$, 95% CI [.808, .840]).

Positive relationship quality items. Twenty adjectives were included for the item pool to develop the PN-RQ positive quality scale. These items were selected to sample the three dimensions of the Semantic Differential (evaluation, potency, activity): *interesting, full, sturdy, enjoyable, good, friendly, hopeful, hot, active, dynamic, deep, fun, pleasant, cheerful, passionate, strong, exciting, alive, energizing*, and *invigorating*. Respondents were instructed to complete these items "*considering only the positive qualities of your relationship, and ignoring the negative ones, evaluate your relationship on the following qualities*."

Negative relationship quality items. In an analogous manner, 20 adjectives were included for the item pool to develop the PN-RQ negative quality scale, reflecting the three Semantic Differential dimensions: *fragile, bad, lonely, static, discouraging, boring, empty, miserable, cold, passive, shallow, tedious, unpleasant, gloomy, distant, weak, dull, lifeless, draining, mind-numbing*. Respondents were instructed to complete these items "*considering only the negative qualities of your relationship, and ignoring the positive ones, evaluate your relationship on the following qualities*."

Results and Discussion

Structure of relationship quality items. To examine the proposed underlying correlational structure of the quality items, we ran CFA models in the Study 1 data using Mplus 7.11 (Muthén & Muthén, 1998–2012). We utilized five widely used fit indices to determine the acceptability of model fit (Kline, 2010): (a) the model chi-squared statistic (the primary index of absolute model fit), (b) the standardized root-mean-square residual (SRMR; values less than .08 suggest acceptable fit), (c) Bentler's comparative fit index (CFI; values above .90 suggesting acceptable fit), (d) the Tucker–Lewis Index (TLI; values above .90 suggest acceptable fit), and (e) the root-mean-square error of approximation (RMSEA; values less than .07 are indicative of acceptable fit). Missing data was negligible in the relationship quality items (0.29%) and so we utilized full information maximum likelihood estimation to handle the small fraction of missing values. Across all of the CFA models, the items were treated as continuous indicators.

Consistent with Osgood's (1964) original work, we first evaluated a hierarchical model in which the items were split into three distinct sets of positive and negative items (evaluation items, activity items, and potency items; see Table 1). We allowed the individual positive and negative lower order factors to form separate positive and negative higher order latent variables. The model demonstrated acceptable fit, $\chi^2(728) = 6960$, $p < .001$; SRMR = .049; CFI = .909; TLI = .903; RMSEA = .069, 95% CI [.067, .070].[1] In contrast, a model in which the 20 positive items loaded on a global positive factor and the 20 negative items loaded on a global negative factor demonstrated poor fit, $\chi^2(735) = 10{,}330$, $p < .001$; SRMR = .054; CFI = .860; TLI = .852; RMSEA = .085, 95% CI [.083, .086].[2] Similarly, a model in which all 40 items loaded on a global relationship quality factor demonstrated poor fit, $\chi^2(736) = 26{,}297$, $p < .001$; SRMR = .137; CFI = .627; TLI = .605; RMSEA = .138, 95% CI [.137, .140].

As seen in Table 1, within the hierarchical CFA model, all of the items significantly loaded on their respective factors and 37 of the 40 items demonstrated notably strong loadings (standardized coefficients ranging from .625 to .929). The lower order latent factors, in turn, all loaded strongly on their corresponding higher order global latent factors (coefficients ranging from .855 to .998). Consistent with this, the CFA estimated that the lower order positive factors correlated from .809 to .908 with one another and the lower order negative factors correlated from .988 to .995 with one another (whereas the correlations between lower order positive and negative factors ranged from −.489 to −.553). Thus, although the CFA results replicated Osgood's evaluation, activity, and potency dimensions in both the positive and negative items examined, their extremely strong correlations further suggested that those six individual dimensions were primarily assessing two dimensions of positive and negative relationship quality. As seen in Table 1, the two higher order latent factors demonstrated a modest correlation ($r = -.577$), supporting their conceptualization as distinct yet related constructs.

Developing positive and negative relationship quality scales. To augment traditional (correlation-based) classical test theory methods (e.g., CFA) with a large-sample probabilistic approach, separate item response theory (IRT; Hambleton et al., 1991) analyses were performed within the Study 1 data on the sets of positive and negative items to identify the items most effective at assessing positive and negative relationship dimensions (shown in Table 1). IRT assumes the items within an analysis are measuring a common construct (i.e., they are unidimensional). Consistent with the higher order CFA findings, EFA analyses on the positive and negative items suggested the items were sufficiently unidimensional for IRT analyses.[3] To perform the

[1] Four-item pairs (always occurring between items in the same lower order factor) demonstrated marked shared variance (suggesting slight overlap/redundancy in item meanings) and so their residual errors were allowed to covary. In addition, one item ("*dynamic*") demonstrated notable cross-loading between both the activity and potency factors, and was therefore allowed to load on both latent variables.

[2] The correlations among the residuals of the 4-item pairs (identified in the hierarchical model) were retained through the two remaining models to allow for more direct comparisons of model fit.

[3] Consistent with the CFA results, an EFA on the 20 positive adjectives in the Study 1 data (principle axis factoring with Oblimin rotation) yielded a dominant first factor accounting for 62% of the variance with a first eigenvalue (12.0) eight times larger than the second eigenvalue (1.5), suggesting that the 20 positive items could be considered a unidimensional pool for IRT (see Reise, Moore, & Haviland, 2010, for a discussion of the appropriateness of such an approach). Similarly, the lower order CFA factors among the negative adjectives demonstrated strong correlations (.809 to .908) and strong factor loadings (.991 to .998). In addition, an EFA on the 20 negative adjectives (PAF with Oblimin rotation) yielded a dominant first factor accounting for 65% of the variance with a first eigenvalue (12.9) over ten times bigger than the second eigenvalue (1.0), suggesting that the 20 negative items could be considered a unidimensional pool for IRT. Given the markedly low rate of missing data across these two sets of items (0.29%), respondents with missing data were dropped from the IRT analyses.

Table 1
Path Coefficients From a Confirmatory Factor Analysis (CFA) in Study 1 (N = 1,814)

Final PN-RQ items		Portion of CFA model: Indicators	β	*SE*	*t*	Final PN-RQ items		Portion of CFA model: Indicators	β	*SE*	*t*
		GLOBAL POSITIVE latent factor						GLOBAL NEGATIVE latent factor			
		POSITIVE ACTIVITY latent factor	.855	.008	105.17			NEGATIVE ACTIVITY latent factor	.991	.004	266.22
		POSITIVE EVALUATION latent factor	.960	.005	186.58			NEGATIVE EVALUATION latent factor	.997	.003	359.12
		POSITIVE POTENCY latent factor	.946	.006	162.69			NEGATIVE POTENCY latent factor	.998	.003	353.48
		POSITIVE EVALUATION latent factor						NEGATIVE EVALUATION latent factor			
P4	P8	enjoyable	.919	.004	212.83		N8	unpleasant	.844	.007	113.58
		good	.914	.005	201.69	N4	N8	miserable	.837	.008	110.68
P4	P8	pleasant	.866	.006	133.56			gloomy	.817	.008	96.73
	P8	fun	.855	.007	123.82	N4	N8	bad	.791	.009	85.05
		cheerful	.854	.007	122.13		N8	dull	.767	.010	75.52
		friendly	.832	.008	106.69			tedious	.753	.011	70.79
		interesting	.772	.010	76.83			lonely	.673	.013	50.60
		POSITIVE POTENCY latent factor						boring	.658	.014	47.64
P4	P8	strong	.861	.007	121.52			NEGATIVE POTENCY latent factor			
	P8	full	.848	.008	112.07	N4	N8	empty	.838	.008	111.03
		sturdy	.832	.008	101.57		N8	weak	.821	.008	100.61
		hopeful	.834	.008	103.08		N8	discouraging	.820	.008	99.52
		deep	.759	.011	70.10			cold	.818	.008	99.03
		dynamic (allowed to cross-load)	.388	.027	14.59			shallow	.708	.012	58.07
		POSITIVE ACTIVITY latent factor						distant	.674	.013	50.60
	P8	energizing	.929	.004	229.22			fragile	.625	.015	42.22
P4	P8	alive	.926	.004	222.84			NEGATIVE ACTIVITY latent factor			
	P8	exciting	.907	.005	185.51			mind-numbing	.822	.009	95.25
		invigorating	.858	.007	124.61	N4	N8	lifeless	.811	.009	92.71
		active	.687	.013	53.05			draining	.782	.010	77.87
		passionate	.427	.020	21.58			static	.721	.012	60.79
		dynamic	.419	.026	15.83			passive	.679	.013	51.24
		hot	.364	.021	17.38			GLOBAL POSITIVE & NEGATIVE correlation	−.577	.017	−33.94

Note. This CFA model was run in Mplus 7.11. The items were treated as continuous indicators using a maximum likelihood estimator, and the model demonstrated acceptable fit, $\chi^2(728) = 6960$, $p < .001$; square-root-mean residual = .049; comparative fit index = .909; Tucker–Lewis Index = .903; root-mean-square error of approximation = .069, 95% confidence interval [.067, .070]. All path coefficients presented were statistically significant at $p < .001$. PN-RQ = Positive Negative Relationship Quality scale; P4 identifies the items of the 4-item PN-RQ positive subscale; P8 identifies the items of the 8-item PN-RQ positive subscale; N4 identifies the items of the 4-item PN-RQ negative subscale; N8 identifies the items of the 8-item PN-RQ negative subscale.

IRT analyses in this study, Graded Response Model (GRM; Samejima, 1997) parameters for the items within each set were estimated with Multilog 7.0 (Thissen, Chen, & Bock, 2002) using marginal maximum likelihood estimation. To assess the quality of the model, we examined residual plots for the item response curves (see Hambleton et al., 1991). As a set, these plots showed evidence of good fit. As described in the introduction, IRT conceptualizes the information provided by an item as that item's ability to discriminate between individuals on the construct being measured (termed θ in IRT equations). Thus an item is considered to be more informative if subjects lower on θ select lower answer choices and subjects higher on θ select higher answers. IRT specifically evaluates how the distributions of the responses for each item map onto the latent θ estimates across all subjects (generating item response curves represented by GRM item parameters) to create information profiles (termed item information curves or IICs) for each item. IICs reveal how much discriminating information each item provides at various levels of θ (ranging from 3 standard deviations below the mean to 3 standard deviations above the mean). IICs therefore synthesize the item parameters estimated by the GRM with IICs of greater height (more information) and greater width (spanning a greater range of θ) identifying highly effective items.[4] Put simply, the greater the area under any IIC, the greater the discriminating information offered by that item, and the precise placement of that curve on the *x*-axis shows the range of θ values across which that item will offer the most information. Although the IICs are not shown in the interest of space, Figures 2A and 2B present test information curves (TICs) that have identical properties to IICs as they are created by summing the IICs of a set of items to model the

[4] The GRM estimates a set of threshold parameters (termed β's or *difficulty parameters* in the IRT equations) that denote the points on the construct of interest (θ) where adjacent answer choices become equally probable for subjects. In addition the GRM estimates one item discrimination parameter (termed α in the IRT equations) that represents how sharp and clean those transitions are between adjacent answer choices. As sharp transitions between answer choices across subjects with different θ levels yield far greater discriminating information for researchers, the height of the IIC for each item is strongly linked to the GRM α estimate for that item. In contrast, the GRM difficulty parameters (β estimates) help to determine where on a range of 3 *SD*s below the mean to 3 *SD*s above the mean each item provides the most information. Thus, the IICs essentially synthesize the item parameters of each item (each item's GRM α and its set of GRM β's) in one graphic profile that can be readily compared with any of the other items in the analysis.

information provided by specific scales. Thus, the IICs (as well as the GRM item discrimination parameters—see footnote 4) were examined to identify the subset of items providing the largest amount of information across the widest range of the trait being measured. This enabled us to identify the 8 (and within those 8, the 4) items most effective at assessing positive and negative relationship dimensions, respectively. Given our conceptual focus of splitting relationship quality into distinguishable negative and positive global evaluations of a relationship, responses to the PN-RQ positive items were summed to create totals representing positive global evaluations of relationships and the responses for the PN-RQ negative items were separately summed to create totals representing negative global evaluations of relationships. Thus, we developed longer 8-item positive ($\alpha = .95$; $\omega = .951$, 95% CI [.946, .956])[5] and 8-item negative ($\alpha = .95$; $\omega = .951$, 95% CI [.945, .957]) versions of the subscales for use when higher levels of precision and power might be required (e.g., in smaller samples). We also created shorter 4-item ($\alpha = .90$; $\omega = .900$, 95% CI [.888, .911]) and 4-item negative ($\alpha = .91$; $\omega = .906$, 95% CI [.893, .918]) versions for use when survey length is a critical factor (e.g., diary studies, telephone surveys).

Evaluating the distinctiveness of the PN-RQ Scales. As seen in Table 2, the PN-RQ positive and negative scales demonstrated a modest negative correlation with one another, suggesting that they share roughly 25% of their variance. In addition, the PN-RQ scales demonstrated moderate associations with a global measure of relationship satisfaction, the CSI-4, suggesting that the positive and negative relationship quality scales shared 42% and 27% of variance with that global measure, respectively. These findings suggested that the PN-RQ scales might offer discriminating information beyond the information provided by the CSI-4.

Discriminatory distinctiveness. As mentioned above, IRT calculates estimates of the latent construct (θ) being analyzed for each subject in the study. Thus, the IRT analyses conducted in the current study provided θ estimates for both positive and negative relationship quality for each of the subjects in the sample. If relationship quality is merely a unidimensional construct, then a unidimensional scale like the CSI-4 should be every bit as effective as the PN-RQ scales at distinguishing groups based on those positive and negative relationship quality θ estimates. To test this, we constructed 10 equally sized groups based on the positive quality θ scores and another 10 equally sized groups based on the negative quality θ scores. We then evaluated the ability of the CSI-4 and the PN-RQ scales to detect differences between adjacent θ groups (e.g., detecting a difference between the 181 respondents with the lowest levels of positive relationship quality and the 181 respondents with next lowest levels of positive relationship quality). As seen in Figures 1A and 1B, both the 8-item and 4-item versions of the PN-RQ scales outperformed the CSI-4 in their abilities to detect adjacent positive or negative relationship quality groups on 15 of the 18 adjacent group contrasts tested (compared as recommended by Meng, Rosenthal, & Rubin, 1992). This suggested that a bivariate conceptualization yielded unique discriminating information on relationship quality by disentangling positive and negative qualities from one another, potentially offering information beyond that provided by an IRT-optimized unidimensional measure of relationship satisfaction (Funk & Rogge, 2007). This represents the first time IRT has been used to help clarify differences between univiariate and bivariate conceptualizations of relationship quality, adding to a growing body of findings (e.g., Mattson et al., 2013). Of course, the preceding interpretation assumes that the CSI and PN-RQ scales are assessing a common underlying construct (relationship quality) from different conceptualizations. An alternative interpretation could be that the CSI and PN-RQ scales are assessing entirely distinct constructs, thereby explaining why the CSI would not be as effective at discriminating in these tests.

Bivariate distinctiveness. To illustrate the immediate advantages of using a bivariate scale of relationship quality, we created median splits using the 8-item versions of the PN-RQ scales. This created 4 distinct groups: *satisfied* respondents (high positive qualities, low negative qualities; $n = 606$), *dissatisfied* respondents (low positive qualities, high negative qualities; $n = 625$), *indifferent* respondents (low positive and negative qualities, $n = 359$), and *ambivalent* respondents (high positive and negative qualities, $n = 215$). We then ran univariate analyses of variance (ANOVAs; followed by Tukey post hoc analyses) to examine differences across these four groups on the relationship process measures included in the study. As seen in Figure 1C, although the CSI-4 was clearly able to distinguish satisfied from dissatisfied groups, its simpler conceptualization failed to uncover differences between respondents with indifferent and ambivalent feelings toward their relationships. However, as seen in Figures 1D, 1E and 1F, the indifferent and ambivalent individuals identified by the PN-RQ scales differed in potentially clinically meaningful ways, with indifferent individuals reporting lower levels of both positive and negative interactions in their relationships as well as lower levels of unforgivingness toward their partners. Taken as a set, the results presented in Figure 1 add to a growing literature on bivariate conceptualizations of relationship quality (e.g., Fincham & Linfield, 1997; Mattson et al., 2013), suggesting that the PN-RQ could offer unique discriminatory insights.

Study 2

Study 2 sought to replicate and extend the findings of Study 1 by (a) sampling a broader range of romantic relationships in a sample of young adults outside of the context of college to ensure that the results would generalize more broadly, (b) including prior two-dimensional scales (the PN-QIMS and the PN-SMD) to enable side by side evaluation of the PN-RQ to existing scales, (c) using a 16-item unidimensional scale (the CSI-16) as the point of comparison to ensure that the PN-RQ continues to offer unique predictive variance beyond that offered by a unidimensional scale of comparable length, and (d) collecting 2-week follow-up data to allow longitudinal validation of the scales and the assessment of their responsiveness to naturally occurring change over 2-weeks.

[5] IRT could be described as a probabilistic approach to measure development in contrast to classic test theory methods that are based on correlational analyses (e.g., EFA, CFA, alpha coefficients). We feel that both approaches offer useful (and often convergent) information. Specifically, we feel that IRT (when used in sufficiently large samples with sufficiently large and diverse item pools) can be a highly effective method of augmenting classic test theory approaches. Thus, although IRT was ultimately used to select the final items, EFAs and CFAs were used to ensure that each set of items being submitted to IRT was sufficiently unidimensional. In a similar vein, although the scales were constructed using IRT, we feel that it is appropriate and helpful to provide alpha and omega coefficients for each scale—providing information on the PN-RQ's internal consistency using a metric that is familiar to other researchers and likely to be used by them in their own samples.

Table 2
Sample Descriptives and Correlations Among Scales

Scale	Range	*M*	*SD*	α	Correlations among constructs 1	2	3	4	5
Study 1 (N = 1,814)									
1. PN-RQ 8-item positive subscale	0–48	39.0	8.0	.95	1				
2. PN-RQ 8-item negative subscale	0–48	5.0	7.9	.95	−.50	1			
3. CSI-4 global satisfaction	0–21	15.8	4.2	.92	**.65**	−.58	1		
4. Positive Interaction	0–8	6.7	1.5	.78	.46	−.33	.39	1	
5. Negative Interaction	0–8	5.8	1.8	.73	−.28	.40	−.29	−.35	1
6. Unforgivingness	0–54	15.0	8.0	.83	−.33	.37	−.41	−.25	.32

					MDC-95		Correlations among constructs										
	Range	*M*	*SD*	α	raw	*SD* units	1	2	3	4	5	6	7	8	9	10	11
Study 2 (N = 787)																	
Positive Relationship Quality Scale																	
1. PN-RQ 8-item positive subscale	0–48	35.9	10.3	.96	10.7	1.04	—										
2. PN-RQ 4-item positive subscale	0–24	18.7	5.0	.94	5.4	1.08	**.96**	—									
3. PN-SMD 7-item positive subscale	0–42	32.7	8.5	.97	8.6	1.01	**.95**	**.94**	—								
4. PNQIMS 3-item positive subscale	0–24	21.1	3.4	.93	4.9	1.44	**.61**	**.65**	**.64**	—							
Negative Relationship Quality Scale																	
5. PN-RQ 8-item negative subscale	0–48	7.4	9.7	.96	9.6	1.01	−.46	−.46	−.49	−.35	—						
6. PN-RQ 4-item negative subscale	0–24	3.6	4.9	.94	4.7	.98	−.47	−.47	−.50	−.36	**.98**	—					
7. PN-SMD 7-item negative subscale	0–42	7.0	8.5	.95	7.3	.87	−.48	−.48	−.51	−.36	**.98**	**.97**	—				
8. PNQIMS 3-item negative subscale	0–24	8.8	6.7	.96	8.9	1.33	−.43	−.43	−.46	−.29	**.69**	**.68**	**.70**	—			
Relationship Functioning Anchor Scale																	
7. CSI-16 global satisfaction	0–81	59.7	16.7	.97	12.0	.72	**.77**	**.77**	**.79**	.56	**−.74**	**−.74**	**−.76**	**−.64**	—		
8. SIRRS Emotional Support	0–56	26.8	16.5	.94			.38	.35	.39	.20	−.39	−.39	−.40	−.33	.46	—	
9. Negative Conflict Behavior	0–63	20.1	14.3	.85			−.18	−.17	−.17	−.15	.37	.36	.37	.38	−.29	−.11	—
10. Unforgivingness	0–54	14.1	8.4	.91			−.37	−.37	−.38	−.32	.42	.43	.43	.48	−.48	−.25	.47

Note. All correlations shown were statistically significant at $p < .001$. Correlations with absolute magnitudes greater than .60 have been bolded for ease of interpretation. Range = lowest and highest possible values based on scoring. MDC-95 = Minimum Detectible Change coefficient (the number of points that scores on a scale much change between two assessment points for that change to be statistically significant for an individual); PN-RQ = Positive Negative Relationship Quality scales; PN-SMD = Positive-Negative Semantic Differential Scales; PNQIMS = Positive-Negative Qualities in Marriage Scale; CSI = Couples Satisfaction Index; SIRRS = Support in Romantic Relationships Scale.

Method

Participants. The sample consisted of 787 respondents in romantic relationships. The participants were predominantly female (66%) and Caucasian (78%) with 6% African American, 5% Latino and 7% Asian and 4% biracial. Respondents averaged 32 years of age (SD = 11.3), 14.5 years of education (SD = 2.2) and 14% of the sample reported having a high school education or less. The respondents reported average yearly incomes of $27,066 ($SD$ = 24,450) with 35% of the sample reporting incomes less than $10,000. A majority of the respondents (44%) were in exclusive dating relationships (together an average of 2.9 years, SD = 3.2), 10% were engaged (together 3.7 years, SD = 2.8), 42% were married (together 12.3 years, SD = 10.2; married 10.0 years, SD = 10.2), and 4% were in nonexclusive dating relationships (together 1.6 years, SD = 2.4).

Procedure. Respondents were recruited from Mechanical Turk to take part in a 15-20min online survey titled, "The Happiness in Relationships Study" and were offered 20 cents of Amazon.com store credit as a recruitment incentive. Respondents were required to be at least 18 years old and currently in a romantic relationship to participate. A total of 745 (95%) respondents provided e-mail addresses to allow us to invite them to the 2-week follow-up assessment. This follow-up assessment contained the same substantive scales of the initial survey along with items assessing global relationship change. Respondents were sent up to three invitation e-mails for the follow-up assessment (at 3-day intervals) and were offered another 20 cents of Amazon.com store credit for completing the follow-up assessment. A total of 473 respondents (60%) completed the follow-up assessment an average of 15.7 days after their initial assessment (SD = 3.8). ANOVA and chi-squared analyses examining attrition showed that the respondents choosing not to provide follow-up data failed to demonstrate any differences from those providing follow-up data on gender, race, rates of employment, or levels of relationship satisfaction as assessed by the CSI-16. However, participants not participating in the follow-up tended to be younger, $F(1, 742) = 19.2$, $p < .001$, $\eta^2 = .026$, with slightly fewer years of education, $F(1, 758) = 17.9$, $p < .001$, $\eta^2 = .024$, and slightly lower annual incomes, $F(1, 742) = 6.7$, $p < .010$, $\eta^2 = .009$.

Measures

Global relationship satisfaction. Respondents completed the 16-item version of the Couples Satisfaction Index (CSI-16; Funk & Rogge, 2007). Responses were summed so higher scores reflected higher levels of satisfaction (α = .97; ω = .976, 95% CI [.972, .979]).

Positive relationship qualities. Participants responded to 3 PN-QIMS positive subscale items, the 7 PN-SMD positive

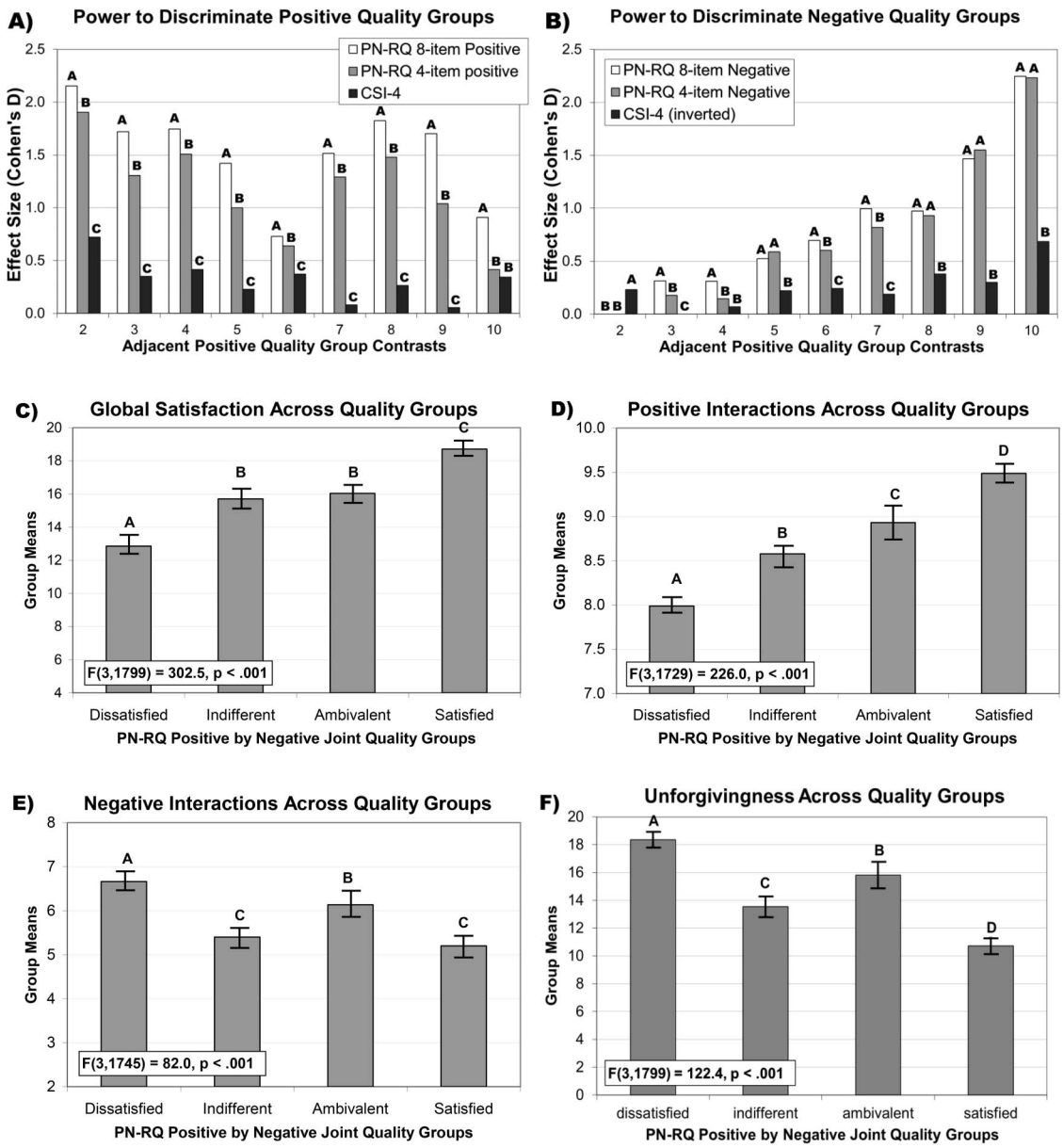

Figure 1. Unique discriminating information provided by Positive–Negative Relationship Quality (PN-RQ) in comparison to a unidimensional scale (the Couples Satisfaction Index; CSI) in Study 1. (Different letters above bars in A and B indicate significantly different effects, $p < .01$, using the strategy detailed by Meng, Rosenthal, & Rubin, 1992. Different letters above bars in C through F indicate significantly different means from Tukey post hoc analyses, $p < .05$).

items (*interesting, full, sturdy, enjoyable, good, friendly, hopeful),* and the 8 PN-RQ positive items developed in Study 1. Responses ranged from *not at all* (0) to *extremely* (8) on the PN-QIMS, from *not at all* (0) to *completely* (6) on the remaining items, and were summed so that higher scores indicated higher levels of positive qualities. On both the PN-RQ and PN-SMD scales, respondents were instructed to complete these items "*considering only the positive qualities of your relationship, and ignoring the negative ones, evaluate your relationship on the following qualities*." The items of PN-QIMS already contained that directive. Responses were summed on these three scales so that higher scores reflected higher positive qualities. The PN-QIMS ($\alpha = .93$; $\omega = .930$, 95% CI [.903, .947]), PN-SMD ($\alpha = .95$; $\omega = .951$, 95% CI [.943, .958]), PN-RQ 8-item ($\alpha = .96$; $\omega = .963$, 95% CI [.956, .968]) and PN-RQ 4-item ($\alpha = .94$; $\omega = .937$, 95% CI [.926, .947]) positive subscales demonstrated high internal consistency in the current sample.

Negative relationship quality items. Participants responded to 3 PN-QIMS negative subscale items (e.g., "considering only the

negative qualities of your partner, *and ignoring the positive ones,* evaluate how negative these qualities are"), the 7 PN-SMD negative items (*fragile, bad, lonely, discouraging, boring, empty, miserable*), and 8 PN-RQ negative items developed in Study 1. On both the PN-RQ and PN-SMD scales, respondents were instructed to complete these items "*considering only the negative qualities of your relationship, and ignoring the positive ones, evaluate your relationship on the following qualities*." All negative subscales were summed so that higher scores indicated higher negative qualities: PN-QIMS (α = .96; ω = .957, 95% CI [.948, .963]), PN-SMD (α = .95; ω = .948, 95% CI [.938, .956]), PN-RQ 8-item (α = .96; ω = .962, 95% CI [.954, .968]) and PN-RQ 4-item (α = .94; ω = .937, 95% CI [.923, .949]) demonstrating high internal consistencies.

Negative conflict behavior. Nine items were used to assess common negative conflict behaviors (e.g., *I swore at my partner, I yelled and screamed at my partner, I have mocked my partner*). Participants rated the frequency of engaging in those behaviors on an 8-point scale (from *never* to *20+ times*) and responses were summed so that higher scores reflected greater amounts of negative interaction (α = .91; ω = .904, 95% CI [.890, .917]).

Emotional support. Eight items from the Support in Romantic Relationships Scale (SIRRS; Dehle, Larsen, & Landers, 2001) were used to assess emotional support. Participants were asked to report how many times (on an 8-point scale, ranging from 0 to 7+) their partners performed eight different behaviors in the past 2 weeks (e.g., *said it was ok to feel the way I was feeling, took my side when discussing my situation*). Responses were summed so that higher scores reflected greater amounts of emotional support (α = .94; ω = .940, 95% CI [.932, .947]).

Unforgivingness. The 9-item Relationship Forgiveness Scale (Fincham et al., 2004) was used to assess individuals' tendencies to be unforgiving following a transgression in a close relationship. Responses were on a 6-point scale (from *strongly disagree* to *strongly agree*) and were summed so that higher scores indicated higher levels of unforgivingness (α = .85; ω = .850, 95% CI [.822, .868]).

Change in relationship quality. Three items were used to assess overall change in relationship quality between the two assessment points in order to identify a "stable" population (respondents perceiving absolutely no change between the two assessment points). These items were prefaced with the instruction, "Since the last survey, how much has your relationship changed (if at all)?" and then asked respondents to rate change on the following items: *feeling close/connected to each other, stability of your relationship, your overall happiness in the relationship*. Respondents rated these items on a 7-point scale (−3 = *has gotten much worse*, −2 = *has gotten somewhat worse*, −1 = *has gotten a little worse*, 0 = *stayed the same*, +1 = *has gotten a little better*, +2 = *has gotten somewhat better*, +3 = *has gotten much better*). Responses to the items were averaged so that higher scores indicated improvement (α = .932; ω = .936, 95% CI [.914, .951]). Eighty individuals had values of zero on this scale, forming a "stable" population when calculating noise in measurement over time (SE_{RM} estimates, see below). Two additional items assessed global change in positive and negative relationship qualities.

Attention/effort. The inconsistency and infrequency subscales of the 33-item Attentive Responding Scale (ARS; Maniaci & Rogge, 2014) were used to screen for excessively inattentive responding. The inconsistency scale is made up of 11 pairs of nearly identical items given at different points in the survey using 5-point response scales (1 = *not at all TRUE* to 5 = *very TRUE*). The scale was scored by summing the absolute differences between responses in each pair of items. Scores exceeding the cutscore of 13.5 were considered excessively inattentive. The infrequency scale was made up of 11 items with such extreme distributions that the vast majority of respondents would provide the same one or two answers. Responses to the items were recoded so that the most probable response had a value of zero and each increasingly unlikely response was worth an additional point. The items were summed and scores exceeding the cutscore of 15.5 were considered excessively inattentive. The 43 (5.5%) individuals identified as excessively inattentive by either scale were omitted from further analyses.

Results and Discussion

Sample descriptives. CSI-16 scores range from 0 to 81 with a global average of 61 (*SD* = 17) for individuals in romantic relationships and a cutscore of 51.5 identifying individuals notably dissatisfied in the relationships (Funk & Rogge, 2007). In comparison to these norms, the sample was modestly happy with mean CSI-16 scores of 60.5 (*SD* = 17.8) in married respondents, 62.9 (*SD* = 14.4) in engaged respondents, 58.5 (*SD* = 16.0) in exclusively dating respondents and 48.1 (*SD* = 11.7) in nonexclusively dating respondents. However, the sample also demonstrated a reasonable range of relationship satisfaction with 28% of the respondents falling below the dissatisfaction threshold.

IRT analysis of positive and negative relationship quality scales. To examine the quality of information provided by the PN-QIMS, PN-SMD and PN-RQ scales, sets of positive and negative quality items were subjected to separate IRT analyses in the Study 2 data. Missing data was negligible in the baseline relationship quality items (0.34%) and so cases with missing values were dropped from the IRT analyses. As mentioned above, TIC's graphically depict the discriminating information offered by a set of items when used as a scale, graphing the information provided by each scale across a wide range of the construct of interest (+/− 3 *SD* around the mean), with greater height on the *y*-axis revealing greater discriminating information. The TICs (Figure 2A) demonstrated that the PN-RQ 8-item positive qualities scale offers more information than the 7-item PN-SMD, effectively operating as a scale roughly 1.48 times longer than the PN-SMD despite having just one more item.[6] Similarly, the PN-RQ 4-item positive scale offers more information than the 3-item PN-QIMS across most of the range of that construct, effectively operating as a scale rougly 4.45 times longer. Similarly, the 8 and 4-item PN-RQ negative qualities scales outperformed the PN-SMD and the PN-QIMS, respectively, across most of the range modeled (Figure 2B).

The TICs further indicated that the positive subscales, as a set, offered lower levels of information at the highest levels of positive

[6] The information curve of a scale can be divided by the information curve of a less informative scale to determine its relative efficacy across a range of θ values (see Hambleton et al., 1991). Those relative efficacy values indicate how much longer the less informative scale would need to be to offer comparable information.

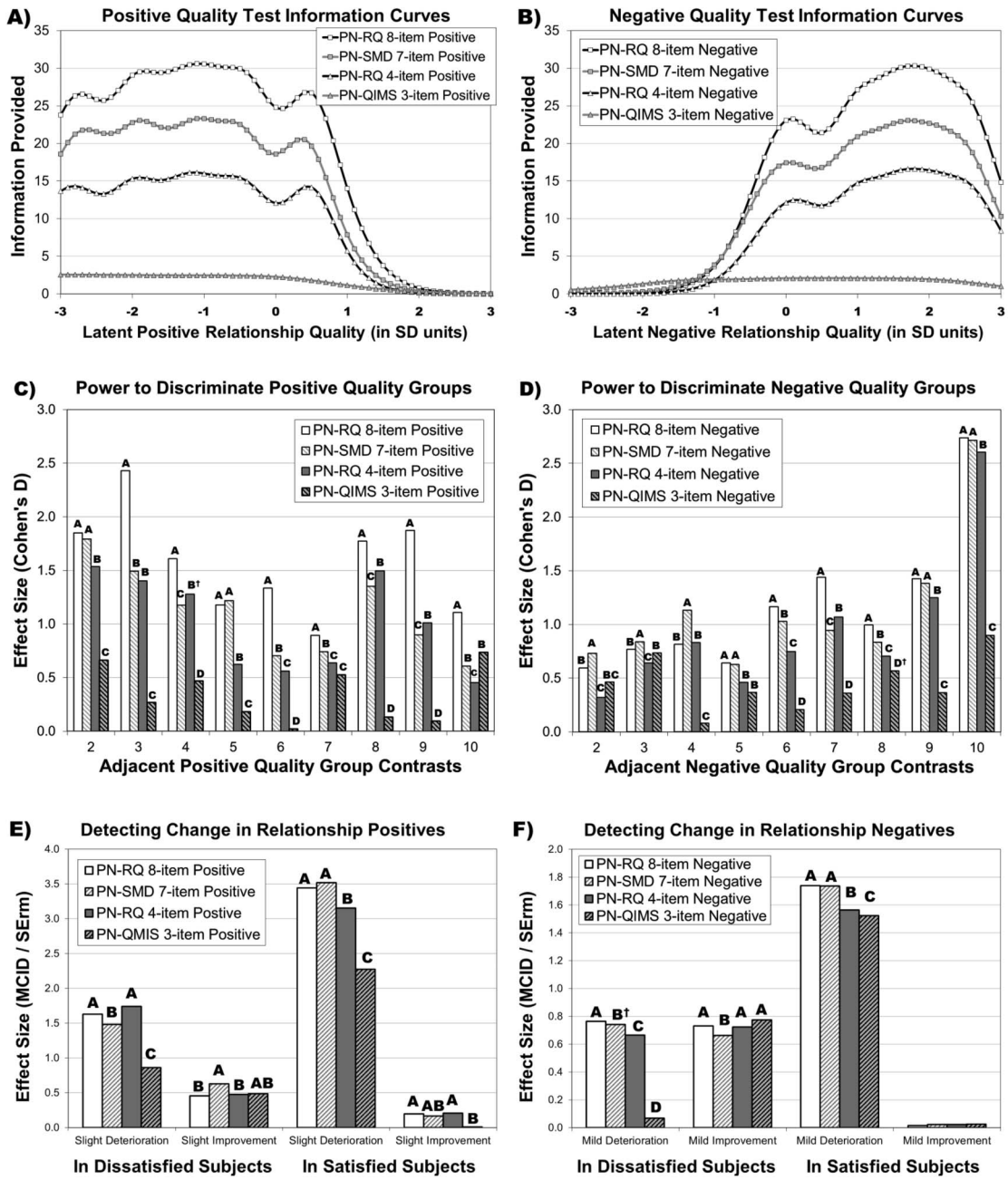

Figure 2. Unique information provided by Positive–Negative Relationship Quality (PN-RQ) in comparison to existing bidimensional scales (Positive and Negative Qualities in Marriage Scale [PN-QIMS], Positive–Negative Semantic Differential [PN-SMD]) in Study 2. (Different letters above histograms indicate effects found to be significantly different at $p < .05$, † at $p < .10$, as assessed with the method suggested by Meng, Rosenthal, & Rubin, 1992).

qualities (Figure 2A). This is a common finding in IRT analyses of positively worded scales (e.g., Funk & Rogge, 2007; Shaw & Rogge, 2016), likely due to a ceiling effect where individuals at very high levels of relationship quality simply select the highest response choices on all the items, making them far harder to distinguish from one another. The TICs for the negative quality subscales (Figure 2B) demonstrated a similar floor effect at the lowest levels of negative quality.

Generalizability of PN-RQ across diverse subsamples. We computed Cronbach's alpha coefficients[5] across a wide range of demographic subsamples to explore how well the PN-RQ subscales function in: females (n = 498), males (n = 246), 18 to 21-year-olds (n = 116), 22 to 30-year-olds (n = 309), 31 to 40-year-olds (n = 179), 41 to 82-year-olds (n = 140), Caucasians (n = 597), African Americans (n = 39), Asians/Pacific Islanders (n = 49), Hispanics/Latinos (n = 51), respondents completing

high school or less (n = 107), some college (n = 320), a bachelor's degree (n = 224), a graduate degree (n = 93), with incomes from 0 to $10,000 ($n$ = 267), $10,001 to $30,000 ($n$ = 206), $30,001 to $50,000 ($n$ = 161), $50,001 or higher ($n$ = 110), who were married (n = 319), engaged (n = 69), or dating (n = 356). The PN-RQ subscales demonstrated high levels of internal consistency across all demographic subsamples tested (α's ranging from .91 to .98) suggesting that the PN-RQ scales will continue to demonstrate excellent levels of internal consistency across a diverse range of future samples.

Precision of positive and negative relationship quality scales. To test if the higher levels of information suggested by the TICs actually provide greater power for detecting differences between groups, we first grouped respondents into 10 equally sized positive quality groups (n's of roughly 71) based on their IRT derived latent positive quality scores (θ estimates) and into 10 comparable negative quality groups based on the θ estimates from the analysis of the negative items. We then calculated the effect sizes (Cohen's d) of each measure for detecting differences between each positive (Figure 2C) and negative (Figure 2D) quality group and the positive or negative group just below it. The 8-item PN-RQ scales performed comparably to and often significantly better than 7-item PN-SMD scales at detecting differences on all 9 of the adjacent positive group contrasts and on 6 of the 9 negative group contrasts (compared as recommended by Meng et al., 1992). Similarly, the 4-item PN-RQ scales performed comparably to and often significantly better than the 3-item PN-QIMS subscales on 8 of the 9 positive group contrasts and on 8 of the 9 negative group contrasts. Thus, when compared with scales of comparable length, the IRT-optimized PN-RQ subscales offered greater power to detect subtle differences in levels of positive and negative relationship quality. Such enhanced precision has been linked to scales offering stronger treatment effects (e.g., Rogge, Crasta, Maniaci, Funk, & Lee, 2016) and is one of the primary benefits of using IRT in scale development. However, by offering evidence to support the enhanced precision of the PN-RQ scales, these result indicate that researchers and clinicians would be able to detect meaningful differences between groups in smaller samples by using the PN-RQ scales.

The results further showed that the 8-item subscales of the PN-RQ outperformed the shorter 4-item versions of those same subscales on 17 of the 18 adjacent group contrasts, suggesting that the longer scales provide greater power for detecting subtle group differences than the shorter scales. Although not surprising, this result helps to highlight the advantages of using the longer version of the PN-RQ when possible. The results presented in Figures 2C and D also build on the findings presented in Study 1, in that the PN-RQ scales not only outperform CSI, but also perform comparably to if not better than the PN-SMD and the PN-QIMS.

Responsiveness to individual change over time. To examine the ability of the scales to detect change over time, we first estimated the noise in measurement over time of the bidimensional scales. Specifically, we estimated the standard error of repeated measurement (SE_{RM}): the distribution of change scores that would be expected in a sample of people experiencing no real change (a stable population). Following the guidelines of Guyatt, Walter, and Norman (1987), we based the estimates of the SE_{RM} on the Mean Squared Error over time (MSE of the within subject effect) from a repeated measures ANOVA on successive scores from 'stable' respondents: SE_{RM} = SQRT(2*MSE). We then used those SE_{RM} estimates to calculate Minimal Detectible Change indices (MDC_{95}; Stratford et al., 1996): $(x_{T1} - x_{T0}) = 1.96^{*} SE_{RM}$, revealing how many points an individual's score must change on a measure between assessments for that change to be statistically significant. Thus, an MDC_{95} of 21 for a scale indicates that an individual's score would need to change at least 21 points between two assessments for that change to be statistically significant for that individual. The MDC_{95} coefficients presented in Table 2 provide future researchers a practical method of grouping individuals in to no change, significant improvement and significant deterioration categories when using any of these scales. This provides critical information to allow researchers to convert treatment effect sizes into a metric more directly relevant to clinical practice as is currently mandated by leading clinical journals like the Journal of Consulting and Clinical Psychology. Thus, in addition to presenting an effect size of .80 for a treatment, the use of MDC_{95} coefficients could convert that into something along the lines of: 75% of individuals receiving the treatment demonstrated significant individual improvement compared with only 30% of individuals in a control condition (see Jacobson & Truax, 1991, for more details on this approach).

As seen in Table 2, the MDC_{95} coefficients for the PN-RQ and the PN-SMD subscales generally indicated that individual scores would need to change by approximately 1 standard deviation on each subscale for that change to be statistically significant whereas individual scores on the PN-QIMS would need to change 1.44 and 1.33 standard deviations on the positive and negative subscales, respectively, to reflect significant individual change. This suggests that the PN-RQ and PN-SMD scales offer researchers instruments that are more responsive to detecting individual change. Although this is not surprising for the PN-RQ 8-item scales and the PN-SMD 7-item scales given their longer lengths, even the PN-RQ 4-item scales seemed to be more sensitive to detecting significant individual change than the PN-QIMS 3-item scales, despite their comparable length. This suggests that cross-sectional precision of the PN-RQ scales translated into high levels of responsiveness to detecting individual change over time.

Responsiveness to change over time at a group level. To determine how effective the various scales were able to distinguish individuals experiencing no change from individuals experiencing small amounts of deterioration or improvement, we estimated Minimal Clinically Important Difference effect sizes (MCID; Guyatt et al., 1987), yielding estimates of how responsive each of the scales might be to naturally occurring change over time. Bigger MCID effect sizes suggest that the corresponding scale shows robust shifts in scores in response to a small amount of change, controlling for the noise over time in that scale. To estimate MCID effects for the scales, we predicted 2-week change scores on each scale in a series of multiple regression analyses using self-reports of global change on the corresponding dimension (either positive or negative relationship qualities) as the primary predictor. This allowed us to determine the average number of points that scores would shift on each scale for each point of global improvement or deterioration reported on the global item asking how subjects' positive or negative relationship qualities have changed in the last 2 weeks, rated on a scale from −3 (*has gotten much worse*) to +3 (*has gotten much better*). To allow for the possibility that the scales might be more responsive to detecting deterioration rather

than improvement, we also included a dichotomous variable coding the direction of change (0 = *improvement*, 1 = *deterioration*). To allow for the possibility that change scores might show greater shifts in respondents with low levels of relationship quality, we included initial levels of the scale being examined as a predictor and a moderator. These analyses yielded separate estimates of the change scores expected on each scale for 1 point of reported deterioration or improvement for couples starting out with either low (−1 *SD*) or high (+1 *SD*) relationship quality. To convert these estimated change scores into MCID effect sizes, we divided them by the SE_{RM} of each scale.

As seen in Figures 2E and 2F, the 8-item PN-RQ subscales produced comparable if not significantly stronger MCID effect sizes than the 7-item PN-SMD subscales on 7 of the 8 effects examined. This suggested that the PN-RQ subscales were comparably responsive to global perceptions of change to the PN-SMD in pre-intervention–post-intervention scores for a single point of improvement or deterioration on the global item asking about overall change. The 4-item PN-RQ subscales yielded stronger MCID effect sizes than the 3-item PN-QIMS on 5 of the 8 effects examined, suggesting that the 4-item PN-RQ subscales were more responsive to detecting mild improvement and deterioration than the PN-QIMS despite being comparable in length. The MCID results further suggested that all of the scales were more effective at detecting deterioration in quality than in detecting improvement, particularly for individuals with the highest levels of relationship quality. Taken as a set, these results suggest that the PN-RQ and PN-SMD subscales are not only better able to detect significant individual change (as assessed by MDC_{95} coefficients) than the PN-QIMS but are also more responsive to global perceptions of change, likely due in part to their longer lengths.

Unique predictive variance offered by the PN-RQ scales. We classified individuals into "significantly improved," "no change," and "significantly deteriorated" groups based on their individual change scores on the CSI-16 and PN-RQ scales (using the MDC_{95} coefficients to assign individuals to categories). As seen in Table 3, the PN-RQ revealed 7 clear change categories in which individuals changed in the same direction on one or both of the subscales, suggesting that positive and negative qualities can change independently over time.

When the CSI-16 change categories (significantly worse, no change, or significantly better) were compared with these 7 PN-RQ change categories, the CSI-16 showed excellent agreement with the joint no-change and the joint significantly worse categories of the PN-RQ (i.e., showing significantly worse scores on both the positive and negative subscales). This suggested that although the CSI and PN-RQ were derived from distinct conceptualizations (unidimensional vs. bidimensional), they would seem to be tapping a common underlying construct of relationship quality. However, the CSI-16 was only able to identify 25–50% of the cases that had been found to be significantly better or worse on just one of the PN-RQ scales. This further helps to underscore the advantages of delineating negative and positive global evaluations with a scale like the PN-RQ, as it offers a method of disentangling how those separate aspects of relationship quality might change independently over time. As a majority of the follow-up respondents demonstrated no significant change on any of the scales (given the relatively short follow-up interval), these results might actually represent an underestimate of the diversity of information that could be obtained using the PN-RQ over a longer interval.

Table 3
Change Groups Revealed by the PN-RQ Over Two Weeks

Type of statistic presented	2-wk Change on PN-RQ Positive Relationship Qualities		
	Sig. worse	No change	Sig. better
2-week change on PN-RQ Negative Relationship Qualities			
Sig. worse			
No. identified by PN-RQ	20	19	
% also identified by CSI	19 (95%)	8 (42%)	
No change			
No. identified by PN-RQ	32	369	6
% also identified by CSI	15 (47%)	336 (91%)	3 (50%)
Sig. better			
No. identified by PN-RQ		16	2
% also identified by CSI		4 (25%)	1 (50%)

Note. Individuals were classified into specific change groups (significantly worse, no change, significantly better) by comparing their change scores on the Positive Negative Relationship Quality scales (PN-RQ) positive and negative subscales to the Minimum Detectible Change coefficient (the number of points that scores on a scale much change between two assessment points for that change to be statistically significant for an individual) for those scales. Individuals were also classified into change groups based on their CSI-16 change scores to examine the degree to which the CSI-16 classification was able to capture the information on outcomes provided by the PN-RQ scales. CSI = Couples Satisfaction Index; Sig. = significantly.

Study 3

The final study sought to extend the findings of Study 2 by directly evaluating the responsiveness of the PN-RQ subscales for detecting pre-intervention–post-intervention treatment effects in a single-group study of the PAIR intervention (Rogge et al., 2013). PAIR promotes relationship health by encouraging couples to use popular media (e.g., movies, TV shows) as a threat-reducing method of facilitating semistructured self-guided discussions of key processes in their own relationships (e.g., support, conflict, forgiveness). Despite the limitations of single-group pre-intervention–post-intervention intervention designs (e.g., lack of a comparison group to control for spontaneous change and placebo effects), Study 3 allowed us to compare the performance of the PN-RQ in comparison to unidimensional relationship quality scales (i.e., the MAT and CSI scales) in a clinical setting.

Method

Participants. Respondents had to be at least 18 years of age and currently in a romantic relationship to participate. The 74 participants providing longitudinal data were 69% female, 88% Caucasian, 4.1% African American, 2.7% Latino, and 5.4% Asian/Pacific Islander. The mean age was 38.9 years (*SD* = 12.6). The average income was $64,428 per year (*SD* = $35,788). Participants reported an average of 16.4 years of education (*SD* = 2.5) with 24.7% having completed less than a bachelor's degree. A majority of the respondents (61%) were married (together for 16.5 years, *SD* = 11.2), with 10% engaged (together for 9.7 years, *SD* = 9.5) and 29% in dating relationships (together for 3.9 years,

SD = 3.9). Most participants (86%) were currently living with their romantic partners, and 35% had children living in the home. The sample was modestly happy with mean MAT scores of 109 (*SD* = 27.4) for dating, 104 (*SD* = 25.1) for married and 128 (*SD* = 17.6) for engaged participants. Using a cut score of 100 on the MAT (e.g., Rogge & Bradbury, 1999), 36% of the married participants, 14% of the engaged participants, and 27% of the dating participants were classified as significantly dissatisfied.

Procedure. Subjects were recruited from the first author's website and heard about the study from: newspaper and magazine articles covering the first author's research (e.g., Parker-Pope, 2014, February 11; 27%), similar TV and radio coverage (e.g., Good Morning America, National Public Radio; 4%), family and friends (35%), Mturk (14%), reddit.com (8%), facebook (7%), and google/yahoo searches (5%). The 20-30min pre and post intervention surveys were given online as were the PAIR materials. Subjects were sent 5 e-mail invitations (spaced 5 days apart) to complete the relationship discussions that formed the core of the PAIR intervention and up to 4 e-mail invitations to complete the post-PAIR survey.

The PAIR program. PAIR took a novel approach by encouraging couples to use popular movies as a method of stimulating discussions about their own relationships. Thus, individuals were encouraged to select 5 movies with their partners from a list of 113 titles (prescreened to ensure that they portrayed sufficient romantic relationship dynamics). This movie list included the 47 titles originally validated for this approach along with 66 newer titles. The couples were specifically encouraged to select a movie, watch it together, and then have a 30–45 min discussion about how their relationship dynamics were similar to or different from the couple on screen. Couples were given semistructured discussion questions to help focus their discussions on key areas of functioning typically covered in skill training workshops (e.g., social support, managing conflict, forgivingness). Feedback from participants suggested that they found this a less-threatening method of having what they found to be productive relationship discussions.

Adherence. To ensure that individuals completed the PAIR discussions, we asked participants to type in brief summaries of what they talked about following each open-ended discussion prompt as they worked their way through the interactive online form. Thus, individuals providing those brief narratives of their conversations following a movie were considered to have completed a PAIR discussion for that movie.

Attrition. A total of 457 participants completed an initial survey and received information on the intervention. Of those respondents, 74 (16%) completed a 1-month follow-up survey. Analyses contrasting the individuals that completed the 1-month follow-up from those that did not suggested that the individuals that provided follow-up data tended to be slightly older, $F(453, 1) = 9.87$, $p < .002$, partial $\eta^2 = .021$. However, chi-square and ANOVA analyses failed to identify any significant attrition differences on gender, race (white vs. all others), income, years of education, relationship stage (dating vs. engaged vs. married), rates of children living in the home, rates of cohabitation, or levels of baseline relationship satisfaction as assessed by any of the measures of relationship satisfaction, suggesting that attrition biases were nominal and that the individuals that completed the follow-up were comparable to the larger sample.

PAIR completers. Of the 74 respondents completing the 1-month follow-up, 54 (73%) engaged in at least one movie-based relationship discussion with their romantic partners (completing 3.5 discussions on average). Analyses suggested that those engaging in movie-prompted discussions tended to have slightly higher levels of education, $F(71, 1) = 5.13$, $p < .027$, partial $\eta^2 = .027$, and were less likely to have children, $\chi^2(1) = 4.78$, $p < .029$, $\varphi = -.259$. However, analyses failed to identify any significant differences on age, gender, race, income, relationship stage, rates of cohabitation, or levels of baseline relationship satisfaction, suggesting that the completers and noncompleters were fairly comparable at baseline.

Measures

Relationship satisfaction. Study 3 included the 15-item MAT (Locke & Wallace, 1959), and the 4-item CSI (Funk & Rogge, 2007). The scales were scored so that higher scores indicated higher levels of relationship satisfaction (CSI-4: $\alpha = .95$, $\omega = .951$, 95% CI [.927, .965]; MAT: $\alpha = .78$, $\omega = .803$, 95% CI [.589, .867]).

PN-RQ. The study included the 8 item (4 positive, 4 negative) version of the PN-RQ using identical instructions and response sets to those described in Studies 1 and 2. The subscales were scored so that higher scores indicated higher levels of positive and negative relationship quality, respectively (PN-RQpos: $\alpha = .94$, $\omega = .936$, 95% CI [.896, .959]; PN-RQneg: $\alpha = .84$, $\omega = .842$, 95% CI [.679, .941]).

Results and Discussion

To determine the effects of PAIR, we calculated pre-intervention–post-intervention treatment effect sizes for each scale.[7] As seen in Figure 3, the MAT and CSI-4 scales showed an improvement in relationship quality over the 1 month of treatment in the 54 participants who completed at least 1 movie discussion. This suggested that PAIR was effective at improving relationship quality. However, the pre-intervention–post-intervention treatment effects for the PN-RQ subscales helped to further clarify this treatment effect. As seen in the graph, the PNRQ showed that the improvements seen on the unidimensional scales (i.e., MAT, CSI-4) were actually a result of a notable drop in negative relationship qualities rather than an increase in positive relationship qualities. Thus, the PN-RQ offered useful clinical insights into an intervention effect beyond what was offered by traditional unidimensional measures of relationship quality.

General Discussion

The results presented offer initial validation of the Positive and Negative Relationship Quality scale (PN-RQ), a measure assessing positive and negative relationship quality as distinct yet related dimensions. The diversity and scale of the samples across the three studies suggest that the PN-RQ is likely to function well across a wide range of populations and across both correlational and intervention studies.

[7] Among these 74 individuals there were no missing values for the relationship quality scales across the two waves of assessment.

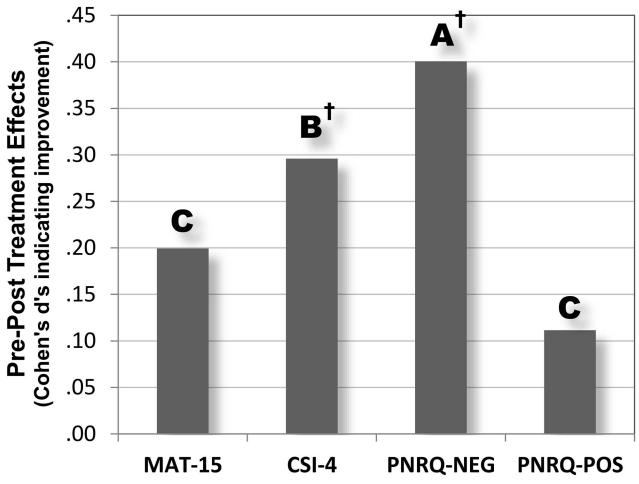

Figure 3. Pre-intervention–post-intervention effect sizes in 54 individuals engaging PAIR in Study 3. Note that different letters suggest significant differences in effect sizes (using the approach suggested in Meng, Rosenthal, & Rubin, 1992 p. 173; $p < .05$, † $p < .10$). For ease of comparison, the direction of the treatment effect for the PN-RQ-negative subscale was reversed so that a higher bar reflects greater improvement.

A Bidimensional Conceptualization Can Provide Greater Insights

Consistent with a growing body of findings in the couples literature (e.g., Fincham & Linfield, 1997; Fincham & Rogge, 2010; Mattson et al., 2007, 2013; Rogge et al., 2013), the current results showed that despite their high levels of IRT-optimized precision, existing unidimensional measures of relationship quality like the CSI might inadvertently obscure meaningful results that could be revealed by the use of a scale like the PN-RQ with its more nuanced bidimensional conceptualization of relationship quality. These results mirror measurement findings in the affect and psychopathology literatures as exemplified by scales like the PANAS and the MASQ which conceptualize positive experiences as related yet distinct from negative experiences. The results also dovetail nicely with a larger body of work suggesting that a diverse array of processes might form more general appetitive and aversive behavioral systems (e.g., Gable et al., 2003). Thus, the current results suggest that marital and dyadic relationship researchers could gain additional discriminating information both cross-sectionally and longitudinally by using a bidimensional scale of relationship quality like the PN-RQ. We assert that bidimensional scales and unidimensional scales fundamentally measure the common construct of *relationship quality* from different conceptual perspectives. Consistent with this, the PN-RQ and CSI scores evidenced moderate convergent validity both in their correlations and in the change they identified across time. However, when directly compared, the bivariate conceptualization (as exemplified by the PN-RQ) seemed to offer unique cross-sectional and longitudinal predictive validity beyond that provided by univariate scales (as exemplified by the CSI). Based on those findings, we would argue that the bidimensional conceptualization offers more nuanced insights into individual differences and change over time that might be obscured by unidimensional conceptualizations of relationship quality. However, it could alternatively be argued that scales like the CSI simply measure a conceptually distinct construct from those being measured by the PN-RQ, thereby explaining the unique predictive variance demonstrated for the PN-RQ over the CSI. Future research across a variety of contexts will help to determine the true discriminant and convergent validity of the constructs assessed by the CSI and PN-RQ scales.

The PN-RQ Is an Optimized Measure

The results of the three studies further suggested that the PN-RQ represents a psychometrically optimized measure of relationship quality. Although the 8-item PN-RQ scales were the longest scales assessed and could therefore be expected to provide greater discriminating information and responsiveness to change, those scales often yielded much stronger effects than comparable scales, suggesting a notably higher level of precision. This optimization of the PN-RQ began with a clear and focused conceptual approach to developing the item pool, heavily informed by Osgood's (1964) work on positive and negative adjectives. The use of more traditional statistical techniques like CFA in combination with less common, more advanced techniques like IRT further helped to identify a highly effective set of items for the PN-RQ. Taken as a set the results suggested, the PN-RQ has the potential to offer researchers greater power for detecting meaningful group differences and treatment effects, a critically important aspect of any scale—particularly in the smaller samples typically associated with treatment studies. This is consistent with recent findings that IRT (when used in large samples with large and diverse item pools) can offer a powerful method of optimizing measures (e.g., Funk & Rogge, 2007; Fraley et al., 2000).

Limitations and Future Directions

The current results need to be viewed in the context of several limitations. First, the studies made use of only self-reported data. Future work could deepen our understanding of the new dimensions of positive and negative relationship quality by linking them to objectively coded behavior within dyadic interactions. Second, the data reported in these studies represents that of only one partner in each relationship. Future work could reveal additional advantages of a bidimensional conceptualization as exemplified by the PN-RQ by collecting dyadic data and directly examining the interplay of perceived relationship qualities between partners across time. Third, Studies 2 and 3 made use of fairly short (1 week and 1 month, respectively) follow-up intervals. Future work should seek to extend these results over longer intervals (e.g., 1–4 years) as that would allow for far more variability in relationship outcomes to be modeled.

Conclusion

Notwithstanding the above limitations, the current studies offer promising support for the PN-RQ scales suggesting that they have the potential to advance both basic and applied research. They not only provide information that beyond that obtained from unidimensional relationship quality scales that dominate marital and close relationship research but they also provide highly precise measures. Finally, their demonstrated ability to detect change, even over relatively short periods, makes them well suited for use in intervention research.

References

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7,* 309–319. http://dx.doi.org/10.1037/1040-3590.7.3.309

Dehle, C., Larsen, D., & Landers, J. E. (2001). Social support in marriage. *The American Journal of Family Therapy, 29,* 307–324. http://dx.doi.org/10.1080/01926180126500

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology, 105,* 399–412. http://dx.doi.org/10.1111/bjop.12046

Fincham, F. D., Beach, S. R., & Davila, J. (2004). Forgiveness and conflict resolution in marriage. *Journal of Family Psychology, 18,* 72–81. http://dx.doi.org/10.1037/0893-3200.18.1.72

Fincham, F. D., & Bradbury, T. N. (1987). The assessment of marital quality: A reevaluation. *Journal of Marriage and the Family, 49,* 797–809. http://dx.doi.org/10.2307/351973

Fincham, F. D., Cui, M., Braithwaite, S., & Pasley, K. (2008). Attitudes toward intimate partner violence in dating relationships. *Psychological Assessment, 20,* 260–269. http://dx.doi.org/10.1037/1040-3590.20.3.260

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. http://dx.doi.org/10.1037/0893-3200.11.4.489-502

Fincham, F. D., & Rogge, R. D. (2010). Understanding relationship quality: Theoretical challenges and new tools for assessment. *Journal of Family Theory and Review, 2,* 227–242. http://dx.doi.org/10.1111/j.1756-2589.2010.00059.x

Fincham, F. D., Rogge, R. D., & Beach, S. R. H. (in press). Relationship satisfaction. In D. Perlman & A. L. Vangelisti (Eds.), *Cambridge handbook of personal relationships*. Cambridge, UK: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511606632.032

Fraley, R. C., Waller, N. G., & Brennan, K. A. (2000). An item response theory analysis of self-report measures of adult attachment. *Journal of Personality and Social Psychology, 78,* 350–365. http://dx.doi.org/10.1037/0022-3514.78.2.350

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. http://dx.doi.org/10.1037/0893-3200.21.4.572

Gable, S. L., Reis, H. T., & Elliot, A. J. (2003). Evidence for bivariate systems: An empirical test of appetition and aversion across domains. *Journal of Research in Personality, 37,* 349–372. http://dx.doi.org/10.1016/S0092-6566(02)00580-9

Guyatt, G., Walter, S., & Norman, G. (1987). Measuring change over time: Assessing the usefulness of evaluative instruments. *Journal of Chronic Diseases, 40,* 171–178. http://dx.doi.org/10.1016/0021-9681(87)90069-5

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

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

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59,* 12–19. http://dx.doi.org/10.1037/0022-006X.59.1.12

Karney, B. R., & Bradbury, T. N. (1995). The longitudinal course of marital quality and stability: A review of theory, method, and research. *Psychological Bulletin, 118,* 3–34. http://dx.doi.org/10.1037/0033-2909.118.1.3

Kline, R. B. (2010). *Principles and practice of Structural Equation Modeling* (3rd ed.). New York, NY: Guilford Press.

Locke, H. J., & Wallace, K. M. (1959). Short marital adjustment and prediction tests: Their reliability and validity. *Marriage and Family Living, 21,* 251–255. http://dx.doi.org/10.2307/348022

Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality, 48,* 61–83. http://dx.doi.org/10.1016/j.jrp.2013.09.008

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. http://dx.doi.org/10.1037/1040-3590.19.1.146

Mattson, R. E., Rogge, R. D., Johnson, M. D., Davidson, E. K. B., & Fincham, F. D. (2013). The positive and negative semantic dimensions of relationship satisfaction. *Personal Relationships, 20,* 328–355. http://dx.doi.org/10.1111/j.1475-6811.2012.01412.x

Meng, X. L., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin, 111,* 172–175. http://dx.doi.org/10.1037/0033-2909.111.1.172

Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus User's Guide. Seventh Edition*. Los Angeles, CA: Author.

Osgood, C. E. (1964). Semantic differential technique in the comparative study of cultures. *American Anthropologist, 66,* 171–200. http://dx.doi.org/10.1525/aa.1964.66.3.02a00880

Parker-Pope, T. (2014, February 11). 'Chick flicks' as couple therapy (p. D4). *The New York Times*.

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment, 92,* 544–559. http://dx.doi.org/10.1080/00223891.2010.496477

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the GLB: Comments on Sijtsma. *Psychometrika, 74,* 145–154. http://dx.doi.org/10.1007/s11336-008-9102-z

Rogge, R. D., & Bradbury, T. N. (1999). Till violence does us part: The differing roles of communication and aggression in predicting adverse marital outcomes. *Journal of Consulting and Clinical Psychology, 67,* 340–351. http://dx.doi.org/10.1037/0022-006X.67.3.340

Rogge, R. D., Cobb, R. J., Lawrence, E., Johnson, M. D., & Bradbury, T. N. (2013). Is skills training necessary for the primary prevention of marital distress and dissolution? A 3-year experimental study of three interventions. *Journal of Consulting and Clinical Psychology, 81,* 949–961. http://dx.doi.org/10.1037/a0034209

Rogge, R. D., Crasta, D., Maniaci, M. R., Funk, J. L., & Lee, S. (2016). *How well can we detect shifts in relationship satisfaction over time? Evaluating responsiveness to change in relationship satisfaction scales*. [Manuscript submitted for publication].

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, NY: Springer. http://dx.doi.org/10.1007/978-1-4757-2691-6_5

Shaw, A. M., & Rogge, R. D. (2016). Evaluating and refining the construct of Sexual Quality With Item Response Theory: Development of the Quality of Sex Inventory. *Archives of Sexual Behavior, 45,* 249–270. http://dx.doi.org/10.1007/s10508-015-0650-x

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. http://dx.doi.org/10.2307/350547

Stanley, S. M., & Markman, H. J. (1997). *The communication danger signs scale*. Unpublished manuscript, University of Denver, Denver, Colorado.

Stratford, P. W., Finch, E., Solomon, P., Binkley, J., Gill, C., & Moreland, J. (1996). Using the Roland–Morris Questionnaire to make decisions about individual patients. *Physiotherapy Canada Physiotherapie Canada, 48,* 107–110.

Thissen, D., Chen, W. H., & Bock, D. (2002). *Multilog user's guide: Multiple, categorical item and test scoring using item response theory* (Version 7.0) [Computer software]. Lincolnwood, IL: Scientific Software International.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology, 54,* 1063–1070. http://dx.doi.org/10.1037/0022-3514.54.6.1063

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. http://dx.doi.org/10.1037/0021-843X.104.1.15

Received January 24, 2016
Revision received August 12, 2016
Accepted August 18, 2016 ■