Journal of Family Psychology
2011, Vol. 25, No. 1, 39–48

0893-3200/11/$12.00 DOI: 10.1037/a0022441

The Reliability of Relationship Satisfaction: A Reliability Generalization Meta-Analysis

James M. Graham, Kate J. Diebels, and Zoe B. Barnow
Western Washington University

We conducted a reliability-generalization meta-analysis of 7 of the most frequently used measures of relationship satisfaction: the Locke–Wallace Marital Adjustment Test (LWMAT), the Kansas Marital Satisfaction Scale (KMS), the Quality of Marriage Index, the Relationship Assessment Scale, the Marital Opinion Questionnaire, Karney and Bradbury's (1997) semantic differential scale, and the Couples Satisfaction Index. Six hundred thirty-nine reliability coefficients from 398 articles and 636,806 individuals provided internal consistency reliability estimates for this meta-analysis. We present the average score reliabilities for each measure, characterize the variance in score reliabilities across studies, and consider sample and study characteristics that are predictive of score reliability. Overall, the KMS and the LWMAT appear to be the strongest and weakest measures, respectively, from a reliability perspective. We discuss the importance of considering reliability invariance when making cross-group comparisons and provide recommendations for researchers when electing a measure of relationship satisfaction.

Keywords: relationship satisfaction, marital quality, reliability, meta-analysis

Supplemental materials: http://dx.doi.org/10.1037/a0022441.supp

Relationship satisfaction is possibly the most widely studied and influential variable in the study of romantic relationships. Subjective relationship satisfaction is associated with a wide variety of important outcomes, including mental health (e.g., Gove, Hughes, & Style, 1983), physical health (e.g., Weiss & Aved, 1978), and child functioning (e.g., Howes & Markman, 1989). Like many such variables in the social sciences, relationship satisfaction is primarily measured with self-report questionnaires. Many self-report measures of relationship satisfaction have been developed, and it can be difficult for researchers to choose between them when looking for a measure. Additionally, modern relationship research has begun to expand beyond its historical focus on heterosexual married couples to include more diverse types of romantic relationships. As a result, it is important to consider the fact that some existing measures may not be appropriately suited for research with cohabiting, same-sex, and other types of couples.

In the present study, we examine the reliability of scores produced by 7 of the most commonly used measures of relationship satisfaction in a reliability generalization (RG) meta-analysis. We use meta-analytic data to test which measures tend to produce reliable scores across a variety of individual and relationship characteristics. In doing so, we provide useful guidance for researchers seeking to identify an appropriate measure of relationship satisfaction for their own research and highlight several important methodological and substantive issues.

James M. Graham, Kate J. Diebels, and Zoe B. Barnow, Department of Psychology, Western Washington University.

Correspondence concerning this article should be addressed to James M. Graham, Department of Psychology, Western Washington University, 516 High Street, Bellingham, WA 98225-9172. E-mail: jim.graham@wwu.edu

Relationship Satisfaction

A wide variety of terms are used to describe the overall quality of a romantic relationship. Terms such as *marital* (or, more broadly, *relationship*) *satisfaction*, *quality*, *adjustment*, and *happiness* are often used synonymously (Heyman, Sayers, & Bellack, 1994). It can be difficult to differentiate between these constructs, as each of these terms are generally poorly defined in the measurement literature (Vaughn & Baier, 1999), and what definitions that do exist are not theoretically derived (Heyman et al., 1994). The lack of consensus over what each of these terms means is further confounded by the fact that measures of these constructs are highly correlated with one another. As a result, some researchers have posited that relationship adjustment, satisfaction, quality, and happiness are either the same construct or part of a higher order factor (Cohen, 1985; Fincham & Bradbury, 1987). We have decided to focus on simple relationship satisfaction, or one's subjective global evaluation of one's relationship. In part, focusing on simple evaluation allows one to separate subjective evaluation from the predictors and consequences of subjective evaluation (Kur-

dek, 1992; Norton, 1983; Sabatelli, 1988). Because the relations between satisfaction and other variables are often of interest to researchers, it is important to be able to measure them separately. Additionally, we use the term *relationship satisfaction* rather than *marital satisfaction* because the former accurately reflects the diverse array of romantic relationship types that are the subject of study in modern relationship research.

Measures of Relationship Satisfaction

The present study considers 7 of the arguably most commonly used measures of relationship satisfaction, the Locke–Wallace Marital Adjustment Test (LWMAT; Locke & Wallace, 1959), the Kansas Marital Satisfaction Scale (KMS; Schumm, Nichols, Schectman, & Grinsby, 1983), the Quality of Marriage Index (QMI; Norton, 1983), the Relationship Assessment Scale (RAS; Hendrick, 1988), the Marital Opinion Questionnaire (MOQ; Huston & Vangelisti, 1991), Karney and Bradbury's (1997) semantic differential scale (SMD), and the Couples Satisfaction Index (CSI; Funk & Rogge, 2007). We have not included what is perhaps the most widely used measure of relationship quality, the Dyadic Adjustment Scale (DAS; Spanier, 1976), because the reliability of DAS scores has been assessed elsewhere (Graham, Liu, & Jeziorski, 2006).

LWMAT

The LWMAT (Locke & Wallace, 1959) is a 15-item test that uses various response formats, including 7-, 6-, 4-, 3-, and 2-choice response formats. The LWMAT measures marital adjustment across a variety of areas. One item measures global happiness, 8 items address agreement on relationship matters such as finances, and 6 items address thoughts and feelings regarding the marriage and one's spouse. Despite these different areas, principle-component analyses have suggested that the LWMAT is primarily unidimensional; the vast majority of items load on one factor, with some items falling onto a much smaller social expectancy factor (Cross & Sharpley, 1981). Because of the different scoring for items, those items describing relationship happiness account for nearly half of the variance in LWMAT scores (Sabatelli, 1988). As such, the LWMAT is often used as a test of general relationship satisfaction. The LWMAT is highly related to other measures of relationship satisfaction and is able to discriminate between low- and high-functioning couples.

Researchers have roundly criticized the LWMAT for having items that are based on a stereotypical marriage in the 1950s that may be no longer appropriate to modern relationships, or perhaps never were appropriate (Cohen, 1985; Sabatelli, 1988). Donohue and Ryder (1982) have pointed out that the fact that the reliability of LWMAT scores was originally tested on couples with either very strong or very weak relationships may have inflated initial reliability estimates. Furthermore, because the LWMAT uses multiple-response formats, scores are likely to violate the essentially tau-equivalent assumption of Cronbach's alpha (Graham, 2006). As a result, Cronbach's alpha may not be an appropriate estimate of reliability. Although its popularity has declined, the LWMAT continues to see occasional use in the relationship research literature.

KMS

The KMS (Schumm et al., 1983) is a short, three-item direct measure of relationship satisfaction. Respondents describe their satisfaction with their spouse, their marriage, and their relationship with their spouse on a 7-point scale. KMS scores have strong correlations with the DAS (Crane, Middleton, & Bean, 2000) and are more strongly correlated with the DAS Satisfaction subscale than with the other subscales (Schumm et al., 1986). KMS scores are able to discriminate between distressed and nondistressed spouses (Crane et al., 2000; Schumm et al., 1985) and possess good construct validity (Schumm, Crock, Likcani, Akagi, & Bosch, 2008). The KMS's greatest advantage is perhaps its brevity; with only three items, it can be added to a more lengthy survey. Like many measures of relationship satisfaction, the KMS tends to produce negatively skewed scores (Schumm et al., 1983) and correlates highly with social desirability (Schumm et al., 1986).

QMI

The QMI (Norton, 1983) is a six-item measure of global satisfaction. The QMI was developed in part as a response to the DAS, which Norton (1983) described as confounding relationship satisfaction with the determinants and consequences of relationship satisfaction. Respondents reply to 5 questions such as, "We have a good relationship," and "Our marriage is strong," on a 7-point scale and one global item, "The degree of happiness, everything considered, in your marriage," on a 10-point scale. The QMI correlates strongly with the DAS (Heyman et al.,1994) and consists of a single factor (Norton, 1983). QMI scores tend to be negatively skewed, and Norton (1983) provided a transformation for normalizing QMI scores.

RAS

The RAS (Hendrick, 1988) is a seven-item questionnaire developed on a sample of dating university students. The RAS is unique in that items are worded so that they are not specific to marriages and are general enough to apply to all types of romantic relationships (Hendrick, 1988). Respondents reply to items such as, "How much do you love your partner?" on a 5-point scale. English- (Hendrick, 1988) and German-language (Sander & Bocker, 1993) versions of the RAS are unidimensional. The RAS correlates highly with the DAS, particularly the Satisfaction subscale (Dinkel & Balck, 2005; Vaughn & Baier, 1999;), and discriminates between dating couples that stay together and those that break up (Vaughn & Baier, 1999). Scores on the RAS tend to be negatively skewed (Dinkel & Balck, 2005).

MOQ

Huston and Vangelisti (1991) originally reported the MOQ as part of a longitudinal study on marriages. The MOQ consists of 10 semantic differential items (e.g., miserable to enjoyable, rewarding to disappointing) and 1 global evaluation item that respondents rate on a 7-point scale adapted from a similar measure of life satisfaction. The semantic differential format of the MOQ allows participants to make overall evaluations of their relationships without reporting on specific processes or behaviors. The MOQ is unidimensional and discriminates between divorced and nondivorced individuals (Huston & Vangelisti, 1991).

SMD

Karney and Bradbury (1997) created a semantic differential measure (SMD) based on a pre-existing measure (Osgood, Suci, & Tannenbaum, 1957). Respondents rate their relationship on 15 adjective pairs (e.g., bad–good, satisfied–dissatisfied, pleasant–unpleasant) using a 7-point scale. The SMD correlates highly with other measures of relationship satisfaction, including the LWMAT, KMS, and QMI. The measure was originally reported as part of a longitudinal study of relationships and has not subsequently seen broad use.

CSI

The CSI (Funk & Rogge, 2007) is a 32-item measure of relationship satisfaction. One global item uses a 7-point scale, whereas the other 21 items use a variety of response anchors, all with 6-point scales. The CSI was developed with a pool of items from a wide variety of measures, including the DAS, LWMAT, KMS, QMI, RAS, and SMD. The CSI represents the only measure of relationship satisfaction examined here developed using item response theory. CSI scores correlate highly with other measures of relationship satisfaction (including all of the measures that initially contributed to its development) and discriminate between distressed and nondistressed relationships (Funk & Rogge, 2007).

Many of the measures described here were initially developed to study marital satisfaction. As such, items of many of these measures ask respondents if they have ever considered divorce and to rate their satisfaction with their spouses, husbands, wives, and marriages. As the field of relationship research has progressed, increasingly more attention has been given to nonheterosexual and nonmarried romantic relationships. In using existing measures with nonmarried couples, it is common for researchers to modify the measures, changing *spouse* to *partner*, *marriage* to *relationship*, and so forth. Researchers vary widely in the extent to which they report these modifications, from explicitly describing the specific changes made to stating nothing and leaving the reader to infer that some changes must have been made given the composition of the sample. Although unavoidable, this practice can have an impact on the psychometric properties of scores produced by a measure. As such, the present meta-analysis provides an important test of how well measures of marital satisfaction translate into measures of relationship satisfaction.

RG

The present study uses an RG (Vacha-Haase, 1998) framework. RG studies are based on several important principles that bear mentioning here. Broadly speaking, reliability coefficients estimate the percentage of variance in a set of observed scores that is due to nonerror factors. The most commonly used measure of reliability, Cronbach's alpha, is a measure of internal consistency reliability and considers the degree of item agreement within a measure. Although other types of reliability, such as interrater and test–retest, are important for measures that rely on observer ratings and measure stable traits, we focus solely on Cronbach's alpha in the present study as it is the most broadly applicable reliability estimate for measures of relationship satisfaction.

Despite the erroneous common parlance of referring to tests as being reliable or unreliable, reliability is a property of test *scores* and not the tests themselves (Thompson & Vacha-Haase, 2000). Put another way, measurements—not measures—are reliable or unreliable. This distinction is quite important, because factors other than the measure itself can influence the reliability of scores produced by a measure. Characteristics of the test taker, and influences of the setting under which the measure was completed, can also influence the reliability of resulting scores.

Rather than reporting the reliability of the data in hand, many researchers instead rely on previous findings of acceptable reliability. That is, researchers might see that the test developers reported acceptable levels of reliability when developing the measure and infer that their own data would also possess acceptable reliability. This process of inferring that the reliability of scores found in one specific instance can be generalized to all possible scores produced by that measure is called *reliability induction* (Vacha-Haase, Kogan, & Thompson, 2000). Reliability induction is problematic because it erroneously presumes that reliability is a static feature of a measure. Because measures can produce scores of differing reliabilities in different circumstances and with different test takers, it is important for researchers to calculate, report, and consider the reliability of the scores used in their research (Wilkinson & the Task Force on Statistical Inference, 1999).

Considering reliability when interpreting substantive results is important for several reasons. Reliability serves as an upper limit on effect sizes (Crocker & Algina, 1986; Worthen, White, Fan, & Sudweeks, 1999). Having reliable scores is important not just in ensuring that one is measuring what one wants to measure but also in determining one's ability to detect the effects of interest. Researchers who rely on reliability induction, when in fact their own scores are less reliable, may underestimate their expected effect sizes when calculating power. Furthermore, when the reliability of scores differs by some status variable, that status variable may appear to moderate the relationship between that vari-

able and another, not because a true moderating relation exists but because of the differential effect of reliability attenuation.

RG meta-analyses serve several important roles. First, they characterize the average score reliability produced by a measure across studies. This provides an important baseline for researchers to consider when choosing measures. Although inductive reasoning can be problematic when inferences are made from a single specific observation, it is strengthened when generalizations are made from a large number of specific instances. Thus, the point estimates of reliability provided by RG studies serve as a stronger basis for reliability induction than those obtained from any single study.

RG meta-analyses also characterize the variability in the reliability of scores produced by a measure across studies. This allows confidence intervals to be constructed around the average reliability of scores produced by the measure, providing an even stronger basis for induction than a single point estimate. If substantial variance in the reliability of scores produced by a measure exists, then RG studies can be used to identify the influence of sample and study characteristics on score reliability. Thus, rather than answering the question, "Is the measure reliable?," RG studies answer the question, "How reliable are the scores produced by the measure across different sample and study characteristics?"

Method

The present meta-analysis considered seven of the most widely used measures of relationship satisfaction: LWMAT, KMS, QMI, RAS, MOQ, SMD, and CSI. We used the Social Sciences Citation Index to identify all the articles that cited any one of the original seven articles that first reported the development of the measures. We conducted a PsycINFO search using terms such as *relationship satisfaction*, *relationship adjustment*, *relationship quality*, *relationship happiness*, and *marital satisfaction* as the search criteria to identify any articles published after the earliest reference (Locke & Wallace, 1959). After discarding duplicates, we identified 1,679 articles for potential inclusion. We then attempted to obtain each of these articles. We discarded 578 articles that did not involve the collection of original data using one or more of the identified measures. We discarded 60 articles because they were not written in English. We were unable to obtain 10 articles through normal interlibrary loan channels.

We then examined the final sample of 1,031 studies in detail. At least two members of the research team coded each article, with any disagreements resolved by the third researcher. The present study only considered Cronbach's alpha and did not include any other type of reliability, such as test–retest or interrater. Of the full 1,031 articles considered for inclusion, 345 (33%) made no mention of the reliability of their scores, 176 (17%) induced that a measure was reliable without reporting a value, 61 (6%) induced a measure was reliable and provided a value from a different study, and 54 (5%) reported the reliability for their data in an unusable form (over .7, between. .8 and .9, etc.) The present study found that 395 (38%) of published studies reported reliability coefficients from their original data collection in a usable form.

For the articles that reported a usable form of a reliability coefficient from their data in hand, we recorded the Cronbach's alphas and a variety of sample characteristics: mean age, relationship length, marital status (coded as percent married), cohabitation status (coded as percent cohabiting), ethnic composition (coded as percent White), sexual orientation (coded as percent heterosexual), proportion of sample recruited from college (coded as percent college sample), gender (coded as percent male), and nationality (coded as percent international). To assess the degree of change in score reliabilities over time, we also noted the publication year of the article. For all articles, we coded sample characteristics at the same level as the reported reliability coefficients. In other words, if the authors reported separate reliability coefficients for men and women, we reported study and sample characteristics separately for men and women. Because of multiple studies per article, we found that a total of 398 articles representing 622 samples and 636,806 individuals provided 639 reliability coefficients for use in the present meta-analysis.

Results and Discussion

To prepare the data for analyses, we conducted a series of transformations. Reliability coefficients such as Cronbach's alpha are often considered r^2-equivalents, as they represent the proportion of measured score variance that is not accounted for by measurement error (Thompson & Vacha-Haase, 2000). We first took the square root of the reliability coefficients to obtain an *r*-equivalent statistic and then applied Fisher's *r*-to-*z* transformation to put them into a format amenable to further computations (Lipsey & Wilson, 2001). We reversed the transformation process on the results when appropriate to report averages and confidence intervals in the original metric of Cronbach's alpha. All analyses were weighted by the inverse variance weight of the reliability coefficients ($n - 3$) and were conducted with Wilson's (2005) random-effects meta-analysis macros for SPSS. In cases where study authors reported reliability coefficients for multiple waves of data from the same sample, we adjusted the weights so that the reliability coefficients were effectively averaged across waves. We could not consistently identify when the same samples provided data for multiple articles. As such, the fact that some samples may have provided multiple reliability coefficients is a necessary limitation of the present meta-analysis.

Table 1 summarizes the descriptive statistics for each of the measures. We used Wilson's (2005) macros to calculate the average reliability coefficient produced by each measure across studies and to calculate 95% confidence intervals about those averages. Table 1 also shows the original reliability coefficients reported during the development of the measures, the highest and lowest reliability coefficients reported in the literature, and the number of reliability coefficients that contributed to the analyses (*k*). We used *Q* tests of homogeneity and I^2 to determine whether the variance in reliability coefficients was different from zero. As

Table 1
Descriptive Statistics for Relationship Satisfaction Score Reliabilities

Measure	Original α	*k*	Mean α	95% confidence interval		Min.	Max.	*Q*	I^2	Orwin's *N*
				Lower	Upper					
LWMAT	NR	122	.785	.770	.799	.42	.93	1,284.6***	91	72
KMS	.98	105	.950	.945	.955	.72	.99	2,479.1***	96	310
QMI	NR	189	.944	.940	.947	.75	.98	3,707.8***	95	524
RAS	.86	196	.872	.863	.881	.68	.98	3,007.4***	94	285
MOQ	.88–.94	19	.921	.906	.934	.83	.96	85.8***	79	43
SMD	.97	3	.976	.967	.982	.97	.98	4.7	57	13
CSI	.98	5	.940	.842	.978	.90	.98	385.1***	99	11

Note. Min. = minimum; Max. = maximum; LWMAT = Locke–Wallace Marital Adjustment Test; NR = not reported; KMS = Kansas Marital Satisfaction Scale; QMI = Quality of Marriage Index; RAS = Relationship Assessment Scale; MOQ = Marital Opinion Questionnaire; SMD = semantic differential scale; CSI = Couple Satisfaction Index.
*** $p \leq .001$.

seen in Table 1, each of the *Q* tests were statistically significant with high heterogeneity coefficients, except for the SMD. The reliability coefficients from measures with statistically significant *Q* statistics and high I^2 coefficients are presumably representing estimates drawn from multiple populations. Because these measures have noteworthy variance in reliability coefficients, it is a logical next step to examine what sample and study characteristics might best explain that variability. It should be noted that the measure without a statistically significant *Q* had only 3 usable reliability coefficients and a medium I^2; the low number of studies was likely responsible for the lack of statistical significance. As such, we cannot make definitive statements about the reliabilities produced by this measure until more data are available. We present the results in Table 1 and describe them further by measure later in this section.

The present study is limited in that it considers only published work, and even then only those articles that reported a useable reliability coefficient could be included. It might be that the reliabilities of scores in studies without reported reliability coefficients were different from those included in the meta-analysis. To assess for publication and reporting bias, we used a derivative of Orwin's (1983) fail-safe *N*. Orwin's *N* is reported in Table 1 and represents the number of studies with a reliability of .5 that would be necessary to bring the average reliability of the measure below .7. As seen in Table 1, the *N*s were relatively high in all cases except for those measures with relatively few available reliability coefficients (SMD and CSI) and suggest that the results we see here are not likely to be the result of publication bias.

We used a variety of approaches to examine the sample and study characteristics that predicted score reliabilities. We used Wilson's (2005) maximum likelihood random-effects regression macro to calculate the regression analyses and correlations. For those measures that provided sufficient experimentwise *k*s, we conducted regression analyses predicting the reliability of test scores with several of the sample characteristics. Ideally, all of the sample characteristics would be entered simultaneously into one regression equation. However, because of inconsistencies in reporting sample demographic information, it was necessary to select a reduced subset of sample characteristics that did not reduce the number of coefficients available for analysis to an untenable level. As such, only 4 measures provided sufficient *k* for this type of analysis, and we had to restrict ourselves to a limited number of predictors. These analyses used maximum likelihood as the method of estimation. We present the results in Table 2 and discuss them by measure later.

We calculated a series of bivariate correlations to test the relations between all of the demographic variables and reliability coefficients. Although this method tends to inflate the experimentwise error rate, it does provide a way to consider all of the data available in this meta-analysis. We

Table 2
Random-Effects Weighted Maximum Likelihood Regression of Score Reliability on Sample Characteristics

Measure	*k*	R^2	Model *Q*	*p*	Residual *Q*	Beta weights				
						White	Male	Married	Heterosexual	Length
LWMAT	35	.160	6.597	.159	34.557	−.125	−.303	.271		.041
KMS	29	.081	2.764	.598	31.497	.204	.012	.180		−.259
QMI	56	.144	9.924	.042	58.883	−.049	−.196	−.033		.356*
RAS	34	.295	12.965	.024	31.039	−.295	−.115	.243	−.185	.254

Note. LWMAT = Locke–Wallace Marital Adjustment Test; KMS = Kansas Marital Satisfaction Scale; QMI = Quality of Marriage Index; RAS = Relationship Assessment Scale.
* $p \leq .05$.

used a pairwise deletion strategy for these analyses. Because each of the correlations has a different associated number of reliability coefficients (*k*), we encourage readers not to over-rely on tests of statistical significance and to instead focus on the relative strength of the associated correlations. Table 3 shows the correlations and their associated *k*s, and we describe them by measure later in this section.

LWMAT

The LWMAT produced scores with relatively low reliability across studies, although the average score reliability of .785 was above Nunnally's (1978) often cited "acceptable for research" value of .7. Overall, the reliability of scores produced by the LWMAT appears to be marginally acceptable and the lowest of the measures examined in the present meta-analysis. This lower reliability may be due to a combination of factors. The LWMAT was originally designed to tap into a variety of factors, including consensus and happiness. Although subsequent research has determined that the LWMAT is primarily unidimensional (Cross & Sharpley, 1981), the presence of a smaller social desirability factor may be in part responsible for the lower reliability seen here. Additionally, the LWMAT uses a variety of response formats. Although items may measure the same construct, the use of different scales violates the assumption of essential tau-equivalence made by Cronbach's alpha. When this assumption is violated, it can cause Cronbach's alpha to underestimate the true reliability of the scores (Graham, 2006). Rather than Cronbach's alpha, a congeneric measure of reliability might be more appropriate.

There was a great deal of variance in LWMAT reliability coefficients across studies. We predicted the reliability coefficients with the average length of the relationship and the percentage of the sample that was White, male, and married. The resulting regression model was statistically nonsignificant, explaining 16% of the variance in LWMAT reliability coefficients. The bivariate correlations suggested that the LWMAT produces more reliable scores with married and heterosexual relationships than with nonmarried and same-sex relationships. Furthermore, LWMAT scores appear to have grown less reliable over time. Although statistically significant, the effects of each of these correlations were relatively small. These correlations support the notion that the LWMAT may be an outdated measure, intended for a specific type of couple that is no longer necessarily the norm in modern relationship research. Although the effects of sample characteristics on the reliability of LWMAT scores are relatively small, when coupled with the relatively low overall reliability, they suggest that the LWMAT is not likely a good first choice when choosing a measure of relationship satisfaction.

KMS

The reliability of KMS scores was quite high, with an average of .95 across studies. This is particularly noteworthy in that the KMS consists of only 3 items, as Cronbach's alpha tends to be higher for longer measures. The fact that a 95% confidence interval constructed around the average KMS reliability was lower than Schumm et al.'s (1983) original estimate suggests that the initial reliability reported during the development of the KMS was likely an overestimate of the true reliability of KMS scores. Thus, researchers inducing the reliability found during the development of the measure are likely misreporting the reliability of their data. Nonetheless, the reliability of the KMS was quite high across studies.

There was a statistically significant amount of variance in KMS reliabilities. We conducted a multiple regression analysis, predicting score reliabilities with the length of the relationship and the percentage of the sample that was White, male, and married. The overall model was statistically nonsignificant, with the predictors explaining only 8% of the variance in reliability coefficients. The bivariate correlations suggest several small but statistically significant effects. Specifically, the KMS tends to produce more reliable scores with unmarried and same-sex relationships than with married and different-sex relationships. A positive trend shows that, since its creation, the KMS has produced somewhat more reliable scores over time. As time has

Table 3
Bivariate Random-Effects Weighted Maximum-Likelihood Correlations (and k Values) Between Score Reliability and Sample Characteristics

Sample characteristics	LWMAT	KMS	QMI	RAS	MOQ
Age	.08 (92)	.16 (82)	.25*** (164)	.42*** (149)	−.39 (16)
Relationship length	.17 (52)	−.20 (50)	.36*** (98)	.35*** (108)	−.43 (14)
% Married	.20* (103)	−.27** (90)	.09 (162)	.36*** (122)	−.04 (17)
% Cohabiting	.13 (103)	−.03 (80)	.14 (132)	.12 (68)	.03 (13)
% White	−.20 (92)	.12 (70)	−.08 (120)	−.17 (120)	−.31 (14)
% Heterosexual	.21* (107)	−.24* (90)	.04 (162)	.04 (121)	
% College	.01 (121)	.02 (101)	−.03 (186)	−.27*** (190)	.18 (19)
% Male	−.15 (121)	.01 (103)	−.15* (183)	−.18* (187)	−.15 (18)
% International	.04 (121)	−.07 (104)	.05 (188)	.21** (196)	
Year	−.17* (122)	.27** (105)	−.04 (189)	.10 (196)	−.45* (19)

Note. LWMAT = Locke–Wallace Marital Adjustment Test; KMS = Kansas Marital Satisfaction Scale; QMI = Quality of Marriage Index; RAS = Relationship Assessment Scale; MOQ = Marital Opinion Questionnaire.
* $p \leq .05$. ** $p \leq .01$. *** $p \leq .001$.

passed, relationship research has begun to focus more and more on diverse types of relationships, beyond traditional heterosexual married couples. Although initially developed as a measure of marital satisfaction, the KMS appears to have crossed the bridge to becoming a measure of relationship satisfaction quite well.

Taken together, these results suggest that the KMS is a strong overall measure of relationship satisfaction. The statistically significant correlations with several sample characteristics are not overly concerning, as the reliability of KMS scores is generally quite high, and the relations with sample characteristics relatively small. These correlations also run contrary to the expected direction for a measure of marital satisfaction; as such, the KMS might be particularly useful for studying same-sex and unmarried couples. Additionally, given the brevity of the KMS, it can easily be added onto other questionnaires or completed on a regular basis as a part of a longitudinal design.

QMI

The reliability of QMI scores was high, with an average of .944 across studies. There was a statistically significant amount of variance in QMI reliabilities. We conducted a multiple regression analysis, predicting score reliabilities with the length of the relationship and the percentage of the sample that was White, male, and married. The overall model was statistically significant, with the predictors explaining 14% of the variance in reliability coefficients. The effect appeared to be largely driven by the effect of relationship length on QMI scores; the QMI tended to produce more reliable scores in older, more established relationships than in newer relationships. The bivariate correlations suggest several noteworthy small to moderate effects. Specifically, the QMI tended to produce more reliable scores with older individuals and with more established relationships. Additionally, QMI scores of women tended to be slightly more reliable than the QMI scores of men, although the effect was quite small.

Overall, the QMI appears to be a strong choice from a reliability perspective. Although the reliability of QMI scores does appear to be stronger in older, more established relationships than in newer relationships, the overall reliability of the QMI is high enough to mitigate some of this. It should be noted that, in using multiple-response formats, the QMI may violate Cronbach's alpha's assumption of possessing an essentially tau-equivalent measurement model (Graham, 2006). As a result, the likelihood of the reliabilities reported here underestimating the true reliability of QMI scores is higher than it would be if a more appropriate congeneric measure of reliability had been used. The fact that the QMI continues to produce reliable scores despite this potential violation suggests strong intrinsic psychometric properties. Although it has a consistent overall reliability, the QMI appears to be a particularly good choice when studying longer term relationships. Because of the possibility of attenuation by reliability, we encourage researchers interested in examining moderators of the relation between relationship length and relationship satisfaction to either use a different measure or to interpret their results with caution.

RAS

The reliability of RAS scores was moderate, with an average of .872 across studies. Subsequent research has shown that the RAS tends to produce more reliable scores than initially indicated during the development of the measure. Thus, researchers inducing reliability from the original study might be underestimating the reliability of their scores. There was a statistically significant amount of variance in RAS reliabilities. We conducted a multiple regression analysis, predicting score reliabilities with the length of the relationship and the percentage of the sample that was White, male, married, and heterosexual. The overall model was statistically significant, with the predictors explaining 30% of the variance in reliability coefficients. Although no individual variables emerged as statistically significant predictors in the context of the other predictors, higher reliabilities were associated with non-White, married, older relationships and, to a lesser extent, women and same-sex relationships. The bivariate correlations suggest several noteworthy effects. Moderate correlations suggest that, despite the fact that it was originally developed on a sample of young dating couples, the RAS produces more reliable scores when administered to older individuals, older relationships, and married couples. Somewhat smaller correlations suggest that the RAS is less reliable with college student versus noncollege student samples, with men versus women, and with U.S. versus international samples.

Overall, the RAS produces scores of an acceptable level of overall reliability, albeit not as strong as those produced by several other measures. Furthermore, the reliability of RAS scores appears to be fairly suspect to sample characteristics. Typically, the psychometric characteristics of scores produced by a measure are maximized for samples similar to those used during the development of the measure. Ironically, a measure that was initially developed with a dating undergraduate sample is best suited (from a reliability standpoint) to studying nonundergraduate, older, well-established marriages. Thus, although the claim of the RAS is that it was intended to be applicable to a wide range of relationship types, it appears to be less applicable than some other measures of relationship satisfaction that were originally developed for use with married couples. Furthermore, given the potential of reliability attenuation, the RAS appears to be particularly ill suited to examining predictors of the differences in relationship satisfaction between different types of relationships.

MOQ

The reliability of MOQ scores was reasonably high and on par with that found in the initial development of the measure, with an average of .921 across studies. Despite the fact that there were relatively few MOQ reliability coefficients available for this meta-analysis, the Q statistic suggests a statistically significant amount of variance in reli-

ability scores across studies. However, there were not a sufficient number of studies available to conduct a multiple regression analysis. Despite the small *k*, the bivariate correlations suggest that the reliability of MOQ scores was highly influenced by sample characteristics. Several moderate correlations suggest that the MOQ produces more reliable scores with younger individuals, newer relationships, and non-White versus White respondents, and it has produced less reliable results over time since the development of the measure. Although only the latter of these emerged as statistically significant, the low number of reliability coefficients available for analysis reduced the available power.

Overall, the MOQ appears to be a fair choice of measure of relationship satisfaction, from a reliability standpoint, although reliability coefficients are not yet available in sufficient numbers for definitive statements to be made. Despite being developed as a measure of marital satisfaction, the MOQ appears to be better suited for young relationships than for longer term relationships. The size of some of the correlations between sample characteristics and MOQ reliability suggests that the MOQ may have problems with reliability attenuation.

SMD

Although the average reliability of the SMD was the highest of the measures of relationship satisfaction examined in the present meta-analysis at .976, the average drew from only 3 reliability coefficients. It is not surprising that the variance between those 3 reliability coefficients was not sufficiently high to result in a statistically significant *Q*. Because of the small *k*, we could not conduct any regression or correlational analyses. The SMD may not be widely used for a number of reasons. Karney and Bradbury (1997) originally published the SMD as part of a longitudinal study, and it was not the subject of a separate psychometric examination. As such, many researchers may not be aware of it. Furthermore, the SMD was developed somewhat recently and competes for researchers' attention with a wide number of more established measures. Although we cannot make any statements about the utility of the SMD with such limited information, the few results that are available indicate that the SMD has the potential to prove to be a reliable and useful measure of relationship satisfaction.

CSI

The average reliability of the CSI was moderately high, with an average Cronbach's alpha of .940; however, only 5 studies contributed to this average. The variance between those 5 reliability coefficients was sufficiently high to result in a statistically significant *Q*. This seems largely due to the fact that the reliability reported during the initial development of the measure (.98) was substantially higher than the reliability reported in subsequent research (.9 to .92). Because of the small *k*, we could not conduct any regression or correlational analyses. The CSI is the most recently developed of relationship satisfaction measures, and it is unsurprising that a number of reliability coefficients sufficient for meta-analysis are not yet available. As the CSI was developed using item response theory, it may yet prove to be a valuable measure beyond more traditional measures of relationship satisfaction. Certainly, the preliminary data appear promising, and the methods used in the development of the measure suggest that it should prove useful across a variety of relationship types.

Summary and General Discussion

There are a wide variety of measures of relationship satisfaction available to romantic relationship researchers. The present study considers 7 of the most common measures and examines the level and sources of influence on the reliability of scores produced by each of the measures. Although reliability is important when selecting a measure, it is not the sole consideration. The specific item content, factor structure, validity, and sensitivity to change of different measures may lend themselves better to some studies than others. Although we make the present recommendations primarily on the basis of reliability, we encourage researchers to consider a wide range of information when choosing a measure of relationship satisfaction.

Of the measures examined here, the KMS appears to be the strongest overall measure on the basis of its reliability. Its brevity and lack of overlap with the determinants and consequences of relationship satisfaction make it even more attractive as a research measure. The QMI also appears to be a good choice for a measure of relationship quality, with a high overall level of reliability. The MOQ and the RAS both appear to be adequate measures of relationship satisfaction from an RG perspective. Both have acceptable overall levels of reliability; however, both appear to be strongly influenced by sample characteristics. The present data suggest that the LWMAT is a poor choice in selecting a measure of relationship satisfaction. Although average scores on the LWMAT are of marginally acceptable reliability, they were the lowest of all those examined here, have worsened over time, and are influenced by sample characteristics. The results of the present meta-analysis suggest that, after over 50 years of service, the LWMAT is best replaced by other available measures. Both the SMD and the CSI appear promising, although more information is necessary before any sort of final determination can be rendered.

In addition to the results pertaining to the individual measures, the present study underscores several other important considerations. The present results suggest the possibility that the construct of relationship satisfaction may be different across the lifespan of a relationship. Across measures, the reliability of relationship satisfaction scores appears to be higher with greater respondent age and relationship length. This may be because the construct of relationship satisfaction is more cohesive in more established relationships, whereas there is more variability between items in young relationships. Future research might use techniques for assessing measurement invariance to determine whether the structure of relationship satisfaction differs across relationships. Alternatively, it may simply be

because younger samples are generally more satisfied with their relationships, resulting in less between-person true reliability and, subsequently, less reliability.

We strongly encourage researchers to always calculate and report the reliability of their own data. The reliability of relationship satisfaction scores can vary widely by sample, and reliability induction can frequently lead to erroneous conclusions. We encourage researchers to be particularly mindful of reliability when making cross-group comparisons. Research on romantic relationships has recently seen an explosion in studies considering different types of relationships. As we test whether, for example, the theories initially developed on married heterosexual couples also apply to same-sex couples, it is important to consider whether variance in reliability coefficients might account for some of the differences (or similarities) we see. Finally, we encourage researchers to be diligent and thoughtful in describing their samples in published research. Science is cumulative, and meta-analysis is one of the primary methods for aggregating information from disparate studies. The quality of meta-analyses is dependent in part on the quality of the initial studies used as data; one way to assist the meta-analytic endeavor is to fully and adequately describe one's study.

Relationship satisfaction is one of the most frequently studied variables in the relationship literature and has been linked to a variety of important processes. As relationship researchers, we have a variety of measures available to us, each with their own strengths and weaknesses. Reliability is by no means the only factor to consider when selecting a measure, but it is an extremely important one, with real potential consequences on substantive findings.

References

Cohen, P. M. (1985). Family measurement techniques. *The American Journal of Family Therapy, 13,* 66–70.

Crane, D. R., Middleton, K. C., & Bean, R. A. (2000). Establishing criterion scores for the Kansas Marital Satisfaction Scale and the Revised Dyadic Adjustment Scale. *The American Journal of Family Therapy, 28,* 53–60.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* New York: Holt, Rinehart, & Winston.

Cross, D. G., & Sharpley, C. F. (1981). The Locke-Wallace Marital Adjustment Test reconsidered: Some psychometric findings as regards its reliability and factorial validity. *Educational and Psychological Measurement, 41,* 1303–1306.

Dinkel, A., & Balck, F. (2005). An evaluation of the German Relationship Assessment Scale. *Swiss Journal of Psychology, 64,* 259–263.

Donohue, K. C., & Ryder, R. G. (1982). A methodological note on marital satisfaction and social variables. *Journal of Marriage and the Family, 44,* 743–747.

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

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.

Gove, W. R., Hughes, M., & Style, C. B. (1983). Does marriage have positive effects on the psychological well-being of the individual? *Journal of Health and Social Behavior, 24,* 122–131.

Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability: What they are and how to use them. *Educational and Psychological Measurement, 66,* 930–944.

Graham, J. M., Liu, Y. J., & Jeziorski, J. L. (2006). The Dyadic Adjustment Scale: A reliability generalizability meta-analysis. *Journal of Marriage and Family, 68,* 701–717.

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 vs. marital adjustment: Construct validity and psychometric properties of three measures. *Journal of Family Psychology, 8,* 432–446.

Howes, P., & Markman, H. J. (1989). Marital quality and child functioning: A longitudinal investigation. *Child Development, 60,* 1044–1051.

Huston, T. L., & Vangelisti, A. L. (1991). Socioemotional behavior and satisfaction in marital relationships: A longitudinal study. *Journal of Personality and Social Psychology, 61,* 721–733.

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.

Kurdek, L. A. (1992). Assumptions vs. standards: The validity of two relationship cognitions in heterosexual and homosexual couples. *Journal of Family Psychology, 6,* 164–170.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis.* Thousand Oaks, CA: Sage.

Locke, H. J., & Wallace, M. (1959). Short marital adjustment and prediction test; Reliability and validity. *Marriage and Family Living, 21,* 251–255.

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

Nunnally, J. C. (1978). *Psychometric theory.* New York: McGraw-Hill.

Orwin, R. G. (1983). A fail-safe *N* for effect size in meta-analysis. *Journal of Educational Statistics, 8,* 157–159.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning.* Urbana: University of Illinois Press.

Sabatelli, R. M. (1988). Measurement issues in marital research: A review and critique of contemporary survey instruments. *Journal of Marriage and the Family, 50,* 891–915.

Sander, J., & Bocker, S. (1993). Die Deutsche Form der Relationship Assessment Scale (RAS): Eine kurze Skala zur Messung der Partnerschaftszufriedenheit [The German form of the Relationship Assessment Scale (RAS): A short scale for measuring satisfaction in a partnership]. *Diagnostica, 39,* 55–62.

Schumm, W. R., Anderson, S. A., Benigas, J. E., McCutchen, M. B., Griffin, C. L., Morris, J. E., & Race, G. S. (1985). Criterion-related validity of the Kansas Marital Satisfaction Scale. *Psychological Reports, 56,* 719–722.

Schumm, W. R., Crock, R. J., Likcani, A., Akagi, C. G., & Bosch, K. R. (2008). Reliability and validity of the Kansas Marital Satisfaction Scale with different response formats in a recent sample of U.S. Army personnel. *Individual Differences Research, 6,* 26–37.

Schumm, W. R., 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.

Schumm, W. R., Paff-Bergen, L. A., Hatch, R. C., Obiorah, F. C., Copeland, J. M., Meens, L. D., & Bugaighis, M. A. (1986). Concurrent and discriminant validity of the Kansas Marital Satisfaction Scale. *Journal of Marriage and the Family, 48,* 381–387.

Spanier, G. B. (1976). Measuring dyadic adjustment: New scales

for assessing the quality of marriage and similar dyads. *Journal of Marriage and Family, 38,* 15–28.

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement, 60,* 174–195.

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement, 58,* 6–20.

Vacha-Haase, T., Kogan, L. R., & Thompson, B. (2000). Sample compositions and variabilities in published studies versus those in test manuals: Validity of score reliability inductions. *Educational and Psychological Measurement, 60,* 509–522.

Vaughn, M. J., & Baier, M. E. M. (1999). Reliability and validity of the Relationship Assessment Scale. *The American Journal of Family Therapy, 27,* 137–147.

Weiss, R. L., & Aved, B. M. (1978). Marital satisfaction and depression as predictors of physical health status. *Journal of Consulting and Clinical Psychology, 46,* 1379–1384.

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604.

Wilson, D. B. (2005). *SPSS for Windows meta-analysis macros* [Computer software]. Available at http://mason.gmu.edu/~dwilsonb/ma.html

Worthen, B. R., White, K. R., Fan, X., & Sudweeks, R. R. (1999). *Measurement and assessment in the schools* (2nd ed.). New York: Longman.

Received October 6, 2010
Revision received November 27, 2010
Accepted November 30, 2010 ■