Personal RELATIONSHIPS

JOURNAL OF THE INTERNATIONAL ASSOCIATION FOR RELATIONSHIP RESEARCH

Personal Relationships, **20** (2013), 422–442. Printed in the United States of America.
Copyright © 2012 IARR; DOI: 10.1111/pere.12001

Assessing relationship quality across cultures: An examination of measurement equivalence

JUDITH GERE AND GEOFF MACDONALD
University of Toronto

Abstract

Researchers are increasingly studying close relationships across cultural contexts. One issue that arises when applying scales originally developed in Western countries to a different cultural context is measurement invariance. Researchers often do not examine whether scales show invariance across cultures and thus can be used with confidence. The goal of this article is to discuss the importance of measurement invariance, to discuss what testing invariance involves, and to test the measurement properties of scales of relationship satisfaction, commitment, intimacy, and trust across 4 samples (United States, Canada, Indonesia, and China). Analyses indicated that weak measurement invariance was met for all 4 scales, and assumptions of strong measurement invariance had to be relaxed for only a few items in each scale. Findings are discussed and recommendations are made regarding using these or other scales that have been shown to meet assumptions of invariance across different cultural groups.

Relationships play a primary role in people's lives and their well-being (Baumeister & Leary, 1995; Gere & MacDonald, 2010). Although research in the area of close relationships has proliferated in recent decades, one important caveat to this work is that the vast majority of data have been collected in the United States (Byrne & Campbell, 1999). In Hofstede's (1980) classic work on the dimensions by which cultures can be classified, the United States is rated as the world's most individualistic nation. Thus, the idea that relationship processes identified in this rather unique cultural context can be expected to generalize to other regions of the world seems to be a highly questionable assumption.

In response to this challenge, many researchers have explored cultural influences on relationships by examining data from multiple cultural groups (e.g., Dion & Dion, 1993; Goodwin & Findlay, 1997; Lalonde, Hynie, Pannu, & Tatla, 2004; Levine, Sato, Hashimoto, & Verma, 1995; MacDonald & Jessica, 2006; Marshall, 2008). Although understanding the role that culture plays in relationship dynamics is very important, it presents specific challenges that are not typically present when researchers are examining only a single cultural group regarding the measurement of relational constructs (Borsboom, 2006; Byrne & Campbell, 1999; Little, 1997). When measurement instruments that were developed and validated with a specific cultural group are later used with a different cultural group, the problem of measurement invariance arises (Borsboom, 2006; Byrne & Campbell, 1999). A scale is said to be measurement invariant when its measurement properties function the same way across different cultural groups—in other words, the items contribute to the construct in question the same way across cultures (French & Finch, 2006; Kline, 2011; Little, 1997).

Judith Gere and Geoff MacDonald, Department of Psychology, University of Toronto.

Judith Gere is presently at the Center for Healthy Aging, The Pennsylvania State University.

This research was supported by a Social Sciences and Humanities Research Council (SSHRC) standard research grant awarded to G.M. and a doctoral SSHRC fellowship awarded to J.G.

Correspondence should be addressed to Judith Gere, The Pennsylvania State University, Center for Healthy Aging, 118 Henderson Bldg., University Park, PA 16802, e-mail: jxg47@psu.edu.

The goal of this article is to discuss the importance of measurement invariance and explain why relationship researchers should attend to this issue when conducting research with multiple cultural groups. As a means of both demonstrating the process involved in achieving invariance and providing useful tools for researchers interested in cross-cultural relationship research, we assess indicators of relationship quality (i.e., intimacy, trust, commitment, and relationship satisfaction) for their measurement equivalence across four cultural groups.

Measurement invariance and its importance

The reliability and validity of many scales used in relationship research have been well established with Western samples, but problems can emerge when researchers conduct research with different cultural groups and use these same scales without a proper examination of the equivalence of the scales across cultures (Borsboom, 2006; Byrne & Campbell, 1999; Little, 1997; Meredith, 1993). The procedure that has been most commonly adopted for cross-cultural studies using self-report scales is back-translation (Byrne & Campbell, 1999). In this procedure, a scale is first translated from the original language to the new language in which it is to be used and then a different individual translates it from the new language back into the original language. Any discrepancies in translation are examined and corrected to preserve the meaning of the original items. The scale is then used to assess the construct of interest in two (or more) cultures.

Although using back-translation is important to ensure proper translation of questionnaire items, the technique on its own does not ensure that the scale items function similarly across cultures. That is, proper translation does not guarantee that the items contribute to the total score on the construct the same way across cultures (Borsboom, 2006; Byrne & Campbell, 1999). For example, imagine that in samples of Chinese and U.S. participants, people from the United States endorse an item on an intimacy scale regarding openly expressing their feelings toward their relationship partner to a greater degree than the Chinese participants. If the higher endorsement of this item by the U.S. participants results not because of greater felt intimacy but because it is less culturally appropriate to openly express feelings in China than in the United States, then the differences in mean levels of intimacy across cultures may be the result of an item that is not equally culturally appropriate. To explicitly test that a scale's items contribute to the overall construct equivalently across cultures, statistical analyses can and should be used.

One area wherein the establishment of measurement equivalence is particularly important is for the examination of mean differences across cultural groups (Borsboom, 2006; Byrne & Campbell, 1999; French & Finch, 2006; Little, 1997; Meredith, 1993; Slof-Op 't Landt et al., 2009). When mean differences across groups are found, these may be either because the groups differ with regards to the construct of interest (the conclusion often inferred by researchers) or because the scale is not culturally invariant and the items contribute differently to the total score on the construct (e.g., different items having different mean scores across cultures). When measurement invariance is not explicitly examined, it is unclear which of these two interpretations is correct. Indeed, invariance could also produce a null effect that hides a meaningful difference across cultures (Borsboom, 2006). Only when measurement equivalence has been established it is safe to conclude that differences between groups (or the lack of differences) are due to the construct being measured (Borsboom, 2006; Byrne & Campbell, 1999; French & Finch, 2006; Little, 1997; Meredith, 1993; Slof-Op 't Landt et al., 2009).

When measurement invariance is explicitly tested, the main issue that is examined is whether the scale being used is functioning the same way across cultural groups (French & Finch, 2006; Kline, 2011; Little, 1997). It is tested statistically with multiple-group confirmatory factor analysis, where the latent factor is the construct being measured (e.g., relationship satisfaction) and the indicators are the individual items on the scale (e.g., rating on the item “I am satisfied with my relationship”; Borsboom, 2006; Byrne & Campbell, 1999; French & Finch, 2006; Little, 1997). Multiple, successively more restrictive models are tested and compared. Researchers begin with a noninvariance model, in which the parameters in the model are allowed to vary freely across the cultural groups (i.e., no restrictions are placed on the model). This model is used as the baseline with which the more restrictive models are compared. When moving to the more restrictive model from the baseline model does not result in a significant decrease in model fit, the more parsimonious (i.e., more restrictive) model is preferred. However, when moving to the more restrictive model results in significant declines in model fit, this indicates that the assumptions of invariance are not met.

In the first restrictive model, the factor loadings of the individual items on the latent factor are examined across cultures for equivalence (i.e., does each item have equal loadings on the satisfaction factor across cultures? Byrne & Campbell, 1999; Little, 1997; Meredith, 1993; Slof-Op ’t Landt et al., 2009). If the factor loadings are equivalent, each item makes an equal contribution to the total score on the construct across cultures, which is the most critical criterion for establishing construct validity (French & Finch, 2006). When the condition of equal factor loadings across cultures for each item is met, the scale meets the criterion for weak measurement equivalence (Meredith, 1993). The establishment of weak measurement invariance is necessary to begin considering any type of cross-cultural comparison (Meredith, 1993).

Second, the intercepts of the items are also examined for equivalence (i.e., does each item have the same intercept across cultures?). If both the factor loadings and the intercepts of the items are equivalent across cultures, strong measurement invariance is established (Meredith, 1993). Finally, the residual variance of the items (i.e., variation in the item scores that is not explained by the construct) may be tested for equivalence in order to establish strict measurement equivalence (Meredith, 1993). This final step, however, is often not conducted, as it is widely recognized that this level of invariance is unrealistic with real data (Borsboom, 2006).

It is important to note that different degrees of measurement invariance are necessary for different types of research questions. Although strong measurement invariance is most desirable, this type of invariance is really only necessary when the goal of the research is to compare mean differences across cultures (Borsboom, 2006; Byrne & Campbell, 1999; French & Finch, 2006; Kline, 2011; Meredith, 1993; Slof-Op ’t Landt et al., 2009). For example, if a researcher would like to compare marital satisfaction in free-choice and arranged marriages and recruits couples from different cultures, it is necessary to conduct such comparisons with scales of marital satisfaction that meet requirements for strong measurement invariance. However, when the primary goal of the researchers is to compare the relations between different constructs across cultures (e.g., testing whether the relation between self-esteem and relationship satisfaction is the same across cultural groups), scales that meet assumptions for weak invariance are sufficient (Borsboom, 2006; French & Finch, 2006; Kline, 2011). It is also possible to have scales that show partial measurement invariance (Kline, 2011). In these scales, only some of the items do not meet the assumptions for strong measurement invariance and these assumptions are relaxed for these particular items only. Partially invariant scales are better than scales with only weak invariance and can also be used to test relations between constructs across cultural groups. Many research questions in the relationship literature focus on comparing the relations between constructs across cultures; thus, weak or partial measurement equivalence is satisfactory for most research questions.

Ideally, statistical examinations of measurement equivalence should be conducted in every study that has as its goal the comparison of cultural groups. However, this may not be feasible in many cases. One of the biggest hurdles that may prevent researchers from being able to test for invariance is sample size (Bentler & Chou, 1987; Kline, 2011). Many researchers work with modest resources that need to be stretched considerably, especially when they are conducting studies that require data collection across multiple cultural groups. This presents a problem because confirmatory factor analysis is feasible only with relatively large samples, with the required sample size depending on the complexity of the scales used and thus, the complexity of the model to be tested with confirmatory factor analysis (Bentler & Chou, 1987; Kline, 2011). The general heuristic is that under normal circumstances, the ratio of sample size to number of free parameters should be no lower than 5:1, but a ratio around 10:1 is preferable (Bentler & Chou, 1987). More complex models have more free parameters and thus need larger sample sizes. When the size of the sample from each culture is small, despite researchers' best intentions, testing for measurement equivalence may not be feasible.

Aspects of relationship quality

In our study, we focused on assessments of the following aspects of relationship quality: intimacy, relationship satisfaction, trust, and commitment. Each of these constructs has received considerable research attention and we chose to focus on them given their established importance in relationship dynamics. Although these aspects of relationship quality are positively correlated with one another, they have also been demonstrated to be distinct constructs, at least in Western contexts (Larzelere & Huston, 1980; Rempel, Holmes, & Zanna, 1985; Sternberg, 1997).

First, we examined intimacy as an important aspect of relationships. Intimacy has been identified as a key element of different types of love and is present in different types of relationships, such as romantic relationships and friendships (Sternberg, 1997). It "refers to feelings of closeness, connectedness, and bondedness" (Sternberg, 1997, p. 315). In general, intimacy represents the level of warmth in a relationship and is indicative of communication, understanding, and sharing within the relationship (Laurenceau, Barrett, & Rovine, 2005; McAdams, 1985; Sternberg, 1997). Intimacy is considered to be a human need or core motive by some researchers (e.g., McAdams, 1985) and has been argued to be one of the greatest rewards of relationships (Laurenceau & Kleinman, 2006). Importantly, intimacy has been linked to many outcome variables as well. For example, intimacy is associated with higher levels of satisfaction with one's relationship (Laurenceau et al., 2005) and higher relationship stability over time (Aron, Aron, & Smollan, 1992; Tsapelas, Aron, & Orbuch, 2009). Some research suggests that couples with higher levels of intimacy allow each other to better fulfill their needs (Prager & Buhrmester, 1998). Thus, intimacy in close relationships has also been associated with higher levels of well-being (Prager & Buhrmester, 1998).

The second relational construct that we examined is relationship satisfaction. This construct is also relevant to different types of relationships and represents an evaluation of the degree to which the relationship meets one's expectations and feelings of positivity about the relationship (Rusbult, 1980, 1983). Being satisfied with a relationship is often studied as a desired outcome, with many researchers focused on identifying factors that influence people's feelings of satisfaction with a particular relationship. There are several reasons why satisfaction is deemed to be an important outcome variable. Satisfaction with a relationship is one of the major factors contributing to the length of the relationship (Rusbult, 1980, 1983). Satisfaction within one's intimate relationship is also one of the strongest predictors of well-being (Heller, Watson, & Ilies, 2004) and has been linked to lower levels of depression (Whitton & Kuryluk, 2012). However, satisfaction is consistently found to decline in marriages over time, setting many relationships up for dissolution (Bradbury, Fincham, & Beach, 2000). Thus, many researchers measure relationship satisfaction as one of the primary indicators of the quality of a relationship and focus on factors that may promote the maintenance of high levels of satisfaction.

The third relational construct we examined is trust. Trust is defined as the belief that a partner can be depended on and will respond reliably, taking into account the needs of the other (Rempel et al., 1985). It has been argued that trust is essential for the development of closeness and intimacy in a relationship and without it, a relationship will not progress (Murray & Holmes, 2009; Rempel et al., 1985). Murray and Holmes (2009) argue that partners must open up to one another in order to allow the relationship to progress to the next level and maintain closeness and intimacy. However, opening up to another involves exposing one's own vulnerabilities and creates the possibility of getting hurt by the other person. Given that exposing one's own vulnerabilities is risky, people are only willing to take the risk of rejection if they trust that the other person will be responsive to their needs. When trust is lost in a relationship, people often decrease their dependence on their partner and distance themselves from their partner (Murray & Holmes, 2009). Loss of trust often results in the breakdown of the relationship, such as when infidelity—a profound betrayal of trust—leads to divorce or dissolution (Snyder, Castellani, & Whisman, 2006). Thus, trust represents an essential ingredient of close relationships.

Finally, we examined relationship commitment. Relationship commitment represents the intention to remain in a relationship, whether out of personal desire or obligation (Rusbult, 1980, 1983). Commitment is particularly important because it is the best known predictor of relationship longevity (Le & Agnew, 2003). Relationship commitment has also been linked to numerous prorelationship behaviors that are important for the maintenance of high-quality relationships. For example, it has been associated with greater forgiveness of partner's transgressions (Finkel, Rusbult, Kumashiro, & Hannon, 2002) and also with willingness to sacrifice in order to benefit the relationship (Van Lange et al., 1997). Thus, high levels of commitment have been identified as being important for both relationship quality and relationship length.

It is important to note that these findings have emerged from research in the United States and hold for participants in Western cultures. It will be important to test whether these findings hold up in other cultures as well when researchers are using scales that show invariance across cultural groups. The tools we aim to provide in this article should be useful in testing these statements in other cultural contexts.

The current study

Our goal in this study was to raise the issue of measurement equivalence and its importance for cross-cultural research, and to demonstrate the process of examining invariance by testing a set of scales of relationship quality for measurement equivalence. If the scales we test meet assumptions for at least weak measurement equivalence across cultures, they can be used in cross-cultural studies with at least some confidence. In this study, we collected data on relationship quality from married adults living in four different nations (United States, Canada, China, and Indonesia). We chose these particular nations to represent differences in individualism and collectivism. Most research has been conducted in the United States where measures have been developed and theories have been proposed and tested. However, the United States is perhaps the world's most individualistic nation (Hofstede, 1980; Oyserman, Coon, & Kemmelmeier, 2002); thus, it is important to test how existing theories hold up in different nations. We also selected Canada as another individualistic nation that is somewhat lower on individualism than the United States (Hofstede, 1980), although these nations have often been treated as about equally individualistic (Oyserman et al., 2002). We chose China and Indonesia to represent nations that are known to be collectivistic (Hofstede, 1980; Oyserman et al., 2002). China is particularly high on collectivism compared to the United States, whereas Indonesia falls somewhere in between the United States and China on collectivism (Oyserman et al., 2002).

We tested the measurement equivalence of self-report scales that assessed intimacy, relationship satisfaction, trust, and relationship commitment. We hope that when it is not feasible to test for measurement equivalence, researchers will at least be able to use scales that have previously been examined and found to be invariant across cultures, such as those included in our study, in order to facilitate proper interpretation of cross-cultural findings in the relationship literature.

Method

Participants and procedures

As part of a larger study, married individuals and/or couples from each of four nations were recruited. In the United States, 109 married individuals (79 women and 30 men) were recruited online through a number of free advertisement websites (e.g., Kijiji). Those who participated were given an option to enter their e-mail address into a drawing for a $50 online gift certificate to Amazon.com. Wives from the United States had an average age of 35.0 years ($SD = 11.62$, range = 19–66) and husbands had an average age of 33.77 years ($SD = 8.39$, range = 22–59). Participants were married for an average of 8.3 years ($SD = 9.05$) and had an average of 1.3 children ($SD = 1.4$). Participants had lived in the United States for an average of 31.7 years ($SD = 12.67$).

In Canada, 50 Canadian married couples ($N = 100$) were recruited for the study through advertisements in local newspapers, bulletin boards, and online. Couples who agreed to participate completed the questionnaires in our laboratory at the University of Toronto, or in public places (e.g., library), or online, and were paid 20 CAD for their participation. Canadian wives had an average age of 48.6 years ($SD = 14.60$, range = 22–81) and husbands had an average age of 49.2 years ($SD = 14.97$, range = 24–82). The couples were married for an average of 20.6 years ($SD = 15.67$) and had an average of 1.6 children ($SD = 1.4$). Participants had lived in Canada for an average of 43.9 years ($SD = 15.08$).

In Indonesia, 50 married couples ($N = 100$) were recruited through a research assistant located on Batam Island, Indonesia. The research assistant contacted couples she knew to request their participation. Participants then spread the information about the study to other couples they knew. The research assistant traveled to each couple's home, where they completed the questionnaires, which were translated into Bahasa Indonesian using the back-translation technique (translated scales are available upon request from the first author). Participants were offered grocery vouchers worth the equivalent of 15 CAD for their participation. Indonesian wives had an average age of 32.2 years ($SD = 8.61$, range = 18–59) and husbands had an average age of 35.4 years ($SD = 8.95$, range = 20–61). Couples were married for an average of 7.5 years ($SD = 8.25$) and had an average of 1.3 children ($SD = 1.14$).

In China, 50 couples ($N = 100$) were recruited through the social network of a faculty member at Chang'an University in the city of Xi'an in Shaanxi province. Couples completed the questionnaires, which were translated into Chinese using the back-translation technique (translated scales are available upon request from the first author), and returned them to the faculty member, who compensated them with 71 yuan (~10 CAD) for their participation. Chinese wives had an average age of 44.8 years ($SD = 14.81$, range = 24–80) and husbands had an average age of 47.1 years ($SD = 15.18$, range = 25–82). Couples were married for an average of 20.9 years ($SD = 15.19$) and had an average of 1.3 children ($SD = 1.13$).

Measures

Intimacy

Intimacy was measured using six items from the intimacy subscale of Sternberg's Triangular Love Scale (Sternberg, 1997). All items were rated on a scale of 1 (*strongly disagree*) to 6 (*strongly agree*). An example item is "I feel emotionally close to my partner" (see Table 1 for all means, standard deviations, and reliability alphas). Only 6 of the 15 intimacy items were administered from the intimacy subscale because of overall study-length concerns (see Table 4 for items).

Satisfaction

Relationship satisfaction was measured using five items (Murray, Holmes & Griffin, 1996a,

Table 1. *Scale alphas, means, standard deviations, and correlations with self-esteem*

	Long version			Short version		
Scales	α	*M* (*SD*)	RSE	α	*M* (*SD*)	RSE
Intimacy						
American sample	.92	4.81 (1.24)	.39	.91	4.78 (1.25)	.40
Canadian sample	.88	5.15 (0.84)	.33	.86	5.17 (0.84)	.35
Indonesian sample	.69	4.86 (0.64)	.12	.63	4.85 (0.65)	.13
Chinese sample	.83	5.05 (0.82)	.37	.82	5.13 (0.80)	.30
Satisfaction						
American sample	.92	4.72 (1.30)	.46	.90	4.61 (1.36)	.44
Canadian sample	.84	5.19 (0.83)	.47	.86	5.13 (0.92)	.47
Indonesian sample	.78	4.94 (0.77)	.18	.78	4.95 (0.78)	.09
Chinese sample	.83	5.21 (0.83)	.41	.85	5.40 (0.77)	.38
Trust						
American sample	.92	4.63 (1.27)	.38	.88	4.65 (1.31)	.37
Canadian sample	.85	5.22 (0.85)	.51	.78	5.20 (0.89)	.49
Indonesian sample	.75	4.96 (0.73)	.11	.53	4.86 (0.79)	.12
Chinese sample	.91	5.25 (0.86)	.35	.84	5.32 (0.85)	.40
Commitment						
American sample	.82	5.34 (1.04)	.21	N/A	N/A	N/A
Canadian sample	.70	4.24 (0.43)	.24	N/A	N/A	N/A
Indonesian sample	.53	4.10 (0.51)	−.06	N/A	N/A	N/A
Chinese sample	.51	4.25 (0.64)	.17	N/A	N/A	N/A

Note. RSE = correlation with self-esteem scores.

1996b), which were rated on a scale of 1 (*not at all true*) to 6 (*extremely true*). An example item is “I am extremely happy with my current romantic relationship” (see Tables 1 and 4).

Trust

Relationship trust was measured using five items that were adapted from the Dyadic Trust Scale (Larzelere & Huston, 1980). All items were rated on a scale of 1 (*not at all true*) to 6 (*extremely true*). An example item is “I feel that I can trust my partner completely” (see Table 1). The full Dyadic Trust Scale was not administered because of length concerns (see Table 4 for items).

Commitment

Relationship commitment was measured using three items (Murray et al., 1996a, 1996b), which were rated on a scale of 1 (*not at all true*) to 6 (*extremely true*). An example item is “I am very committed to maintaining my relationship” (see Tables 1 and 4).

Data analysis

We used the statistical software MPlus 5 (Muthén & Muthén, 2007) to conduct multiple-group confirmatory factor analysis in order to test for measurement invariance. We conducted our analyses for each of the four constructs separately. Each construct was modeled as a latent variable using the individual items of the scale as its observed indicators (see Table 2 for zero-order correlations between the items for each sample). We allowed residual correlations between items for scales that consisted of more than three items, given that this is commonly expected in measures of psychological constructs (Bentler & Chou, 1987; McGrath, 2009). We used the chi-square difference test to compare the fit of the more restrictive models to the baseline model with no cross-group restrictions (Kline,

Table 2. *Zero-order item correlations for each scale by sample*

	Intimacy items					Satisfaction items			
	1	2	3	4	5	1	2	3	4
U.S. sample									
1. Item 1	1.00					1.00			
2. Item 2	.72	1.00				.79	1.00		
3. Item 3	.60	.61	1.00			−.70	−.71	1.00	
4. Item 4	.72	.62	.83	1.00		.77	.73	−.56	1.00
5. Item 5	.62	.61	.57	.55	1.00	.78	.81	−.62	.76
6. Item 6	.77	.70	.74	.74	.75				
Canadian sample									
1. Item 1	1.00					1.00			
2. Item 2	.62	1.00				.85	1.00		
3. Item 3	.54	.41	1.00			−.35	−.41	1.00	
4. Item 4	.64	.64	.60	1.00		.70	.64	−.51	1.00
5. Item 5	.38	.58	.35	.56	1.00	.64	.59	−.32	.56
6. Item 6	.58	.48	.58	.69	.70				
Indonesian sample									
1. Item 1	1.00					1.00			
2. Item 2	.28	1.00				.61	1.00		
3. Item 3	.35	.32	1.00			−.45	−.48	1.00	
4. Item 4	.41	.41	.34	1.00		.61	.43	−.28	1.00
5. Item 5	.30	.32	.35	.09	1.00	.55	.45	−.18	.63
6. Item 6	.28	.22	.17	.09	.34				
Chinese sample									
1. Item 1	1.00					1.00			
2. Item 2	.39	1.00				.81	1.00		
3. Item 3	.34	.37	1.00			−.45	−.39	1.00	
4. Item 4	.46	.29	.56	1.00		.64	.57	−.41	1.00
5. Item 5	.33	.50	.45	.48	1.00	.66	.54	−.48	.67
6. Item 6	.58	.50	.44	.51	.64				

2011). In this test, the chi-square value of the less restrictive model (i.e., the model without the constraints) is subtracted from the chi-square value of the more restrictive model (i.e., the model with the equality constraints imposed) and the resulting chi-square value is tested for significance with the difference in the degrees of freedom between the two models (Kline, 2011). If the difference between the two models is not significant, the more restrictive model with greater degrees of freedom is preferred (i.e., the model with invariance constraints).

We first ran our analysis of each construct without imposing any invariance restrictions across the groups (i.e., estimates were allowed to vary freely across the groups) to get the model fit indices for this baseline, noninvariance model. Next, we held the factor loadings and the intercepts of items on the latent variable constant across the four groups to conduct our test for strong measurement invariance (Muthén & Muthén, 2007). We compared the fit of this model with that of the baseline model. If the decrease in model fit was significant (as indicated by the change in chi-square value), we relied on the model modification indices to identify violations of measurement invariance. Modifications were made one by one, where needed, based on

Table 2. *Continued*

	Trust items				Commitment items	
	1	2	3	4	1	2
U.S. sample						
1. Item 1	1.00				1.00	
2. Item 2	.65	1.00			.84	1.00
3. Item 3	.85	.69	1.00		−.50	−.50
4. Item 4	.65	.69	.64	1.00		
5. Item 5	.82	.69	.79	.66		
6. Item 6						
Canadian sample						
1. Item 1	1.00				1.00	
2. Item 2	.55	1.00			.63	1.00
3. Item 3	.60	.49	1.00		−.35	−.46
4. Item 4	.49	.54	.38	1.00		
5. Item 5	.73	.45	.74	.50		
6. Item 6						
Indonesian sample						
1. Item 1	1.00				1.00	
2. Item 2	.35	1.00			.26	1.00
3. Item 3	.33	.13	1.00		−.22	−.37
4. Item 4	.48	.32	.54	1.00		
5. Item 5	.49	.38	.32	.65		
6. Item 6						
Chinese sample						
1. Item 1	1.00				1.00	
2. Item 2	.66	1.00			.62	1.00
3. Item 3	.64	.57	1.00		−.18	−.18
4. Item 4	.69	.76	.62	1.00		
5. Item 5	.70	.60	.82	.76		
6. Item 6						

large modification indices (i.e., indices of at least 5.5), always starting with the modification that indicated the largest violation of invariance. Thus, we created a model that was partially invariant. We compared the fit of the final partial-invariance model with the baseline, noninvariance model to assess whether it resulted in a decrease in model fit.

We tested mean differences between the cultures on the constructs of interest (using all of the items for each scale). To conduct these analyses, we used the sample from the United States as the comparison group. We constrained the mean of the latent factor representing the construct to be equal to that of the U.S. sample for each other sample, one by one. After each constraint, we tested whether the fit of the model has become significantly worse, using the chi-square values. In this test, the chi-square value of the less restrictive model (i.e., the model without the constraint) is subtracted from the chi-square value of the more restrictive model (i.e., the model with the equality constraint imposed) and the resulting chi-square value is tested for significance with 1 *df.* If the chi-square value is significant, the restrictive model provides a worse fit to the data and the means are not equivalent. We also examined whether the variance of the construct of interest was invariant across cultures. In other words, we tested a model in which we constrained the variance of the latent factor representing the construct to be equal across the groups (latent factor variance invariance model) and compared the fit of this model with the model in which the latent factor variances were not constrained to examine changes in model fit.

We also tested a model in which we included all four constructs (intimacy, satisfaction, trust, and commitment) to test the equivalence of the relations between these constructs across the four samples. First, we tested a baseline model, in which the constructs were allowed to covary with one another freely in each sample. We then restricted the relations between the constructs to be equal across the samples, such that the correlations between all of the constructs were the same in each of the four samples. This model tests the invariance of the covariance of the latent factors across the samples. Finally, we restricted the latent factor variances to be equal across the samples and compared the fit of this latent factor variance invariance model with the baseline model.

In addition to testing measurement properties of the scales we used, when strong invariance assumptions were violated, we attempted to construct a shorter measure that would meet criteria for strong invariance by removing the items that did not meet these assumptions. Using only the items that met assumptions for strong invariance, we again tested a baseline model, where we did not constrain estimates across samples. We then compared the model fit of the strong invariance model with the baseline model to assess whether it results in significant changes in model fit. Finally, we tested the invariance of the latent factor variances by constraining these across samples and comparing the fit of the latent factor variance invariance model to the strong invariance model to assess changes in model fit.

For each of the models, we examined the model fit indices to evaluate the overall fit of the model to the data. To evaluate the fit of the models, we relied on the following criteria: a nonsignificant chi-square value, a comparative fit index (CFI) greater than .90, a root mean square error of approximation (RMSEA) of .08 or less, and a standardized root mean square residual (SRMR) of .10 or less (Kline, 2011). To also account for the nested nature of the data in our samples where couples were the participants (participants nested within couples), we used the cluster command in Mplus to adjust the standard error estimates to account for nonindependence (Muthén & Muthén, 2007).

Results

Intimacy

We first conducted our analysis for the intimacy scale. Table 3 shows the comparisons of the different models for intimacy. Moving from the noninvariance model to the strong measurement-equivalent model resulted in a marginally significant drop in model fit. All of the items showed high loadings on the latent factor and met the conditions of weak measurement equivalence (i.e., equal factor loadings across cultural groups). The majority of the items also met conditions for strong measurement equivalence (i.e., equal intercepts), with the exception of the item “I share deeply personal information about myself with my partner,” for which assumptions of equivalence had to be relaxed for the Chinese sample (see Table 4 for scale items). When the assumption of equivalence was relaxed in the model for the Chinese sample, the resulting partial-invariance model fit the data well, $\chi^2(62) = 84.02$, $p = .033$, CFI = .960, RMSEA = .059, SRMR = .135, and did not result in decreases in model fit from the baseline model. Figure 1 shows the final model for the full intimacy scale with the fully standardized factor loadings. Factor loadings differ for the samples because the equality constraints are imposed on the unstandardized loadings. When this single item was eliminated to create a shorter scale, the remaining items met the criteria for strong measurement equivalence. The fit of this shorter scale was also acceptable, $\chi^2(41) = 52.66$, $p = .105$, CFI = .973, RMSEA = .053, SRMR = .122. Thus, the items in this scale of intimacy seem to function similarly across the different

Table 3. *Comparison of models with different degrees of restriction for each construct*

Model	χ^2 (*df*)	CFI	RMSEA	δ χ^2 (δ *df*)	*p*
Intimacy					
Noninvariance model	72.64 (47)	.954	.074	—	—
Full metric-invariance model	95.03 (63)	.942	.071	22.39 (16)	~.10
Partial metric-invariance model	84.02 (62)	.960	.059	11.38 (15)	>.05
Latent factor variance invariance model	111.34 (65)	.916	.084	27.32 (3)	<.05
Noninvariance model—short scale	43.15 (29)	.967	.070	—	—
Full metric invariance model—short scale	52.66 (41)	.973	.053	9.51 (12)	>.05
Latent factor variance invariance model—short scale	81.65 (44)	.913	.092	28.99 (3)	<.05
Satisfaction					
Noninvariance model	60.03 (28)	.940	.107	—	—
Full metric-invariance model	86.83 (42)	.916	.103	26.80 (14)	<.05
Partial metric-invariance model	63.60 (40)	.956	.077	3.57 (12)	>.05
Latent factor variance invariance model	88.40 (43)	.915	.102	24.80 (3)	<.05
Noninvariance model—short scale	28.54 (6)	.925	.193	—	—
Full metric invariance model—short scale	21.45 (12)	.969	.089	0 (6)	>.05
Latent factor variance invariance model—short scale	37.02 (15)	.927	.121	15.57 (3)	<.05
Trust					
Noninvariance model	116.62 (31)	.872	.166	—	—
Full metric-invariance model	98.97 (43)	.916	.114	0 (12)	>.05
Partial metric-invariance model	63.44 (41)	.966	.074	0 (10)	>.05
Latent factor variance invariance model	80.89 (44)	.945	.091	17.45 (3)	<.05
Noninvariance model—short scale	8.44 (6)	.990	.064	—	—
Full metric invariance model—short scale	10.45 (12)	1.00	.000	2.01 (6)	>.05
Latent factor variance invariance model—short scale	27.78 (15)	.950	.092	17.33 (3)	<.05
Commitment					
Noninvariance model	7.07 (6)	.988	.042	—	—
Full metric-invariance model	19.28 (12)	.918	.078	12.21 (6)	>.05
Latent factor variance invariance model	38.14 (15)	.739	.124	18.86 (3)	<.05
Model including all measures					
Latent factor noninvariance model	1,012.04 (648)	.901	.075	—	—
Latent factor variance invariance model	1,066.16 (660)	.889	.078	54.12 (12)	<.05
Latent factor variance and covariance invariance model	1,103.33 (678)	.884	.079	91.29 (30)	<.05

cultural groups, with most of the items meeting criteria for strong invariance.

Next, we examined mean differences in intimacy between the four cultural groups using the partial-invariance model (see Table 5 for means and difference tests). Given that the majority of work with these scales has involved participants from the United States, we used the U.S. sample as our comparison group. Constraining the means to be equal to the mean of the U.S. sample did not result in significant drops in

Table 4. *Scale items and assumptions of invariance met*

Intimacy items
Items showing strong invariance
I communicate well with my partner.
I feel that I really understand my partner.
I feel that my partner really understands me.
I am willing to share myself and my possessions with my partner.
I feel emotionally close to my partner.
Items showing weak invariance
I share deeply personal information about myself with my partner.
Satisfaction items
Items showing strong invariance
I am extremely happy with my current romantic relationship.
I have a very strong relationship with my partner.
I am perfectly satisfied in my relationship.
Items showing weak invariance
I do *not* feel that my current relationship is successful.
My relationship with my partner is very rewarding (i.e., gratifying and fulfilling).
Trust items
Items showing strong invariance
When we are dealing with an issue that is important to me, I feel confident that my partner will put my feelings first.
I feel that I can trust my partner completely.
My partner is a thoroughly dependable person.
Items showing weak invariance
I feel that my partner can be counted on to help me.
My partner does not hesitate to make significant sacrifices to strengthen our relationship.
Commitment items
Items showing strong invariance
I am very committed to maintaining my relationship.
I have made a firm promise to myself to do everything in my power to make my relationship work.
I do *not* feel any moral duty or obligation to continue my relationship.

model fit (although the Chinese model change was marginal); thus, intimacy scores do not differ across the four samples. Finally, we tested whether the variance of the intimacy scores was equivalent across the samples. When we constrained the variance of the latent intimacy factor to be equivalent across the groups, the decrease in model fit was significant, indicating that the variance of intimacy scores differs across the groups.

Relationship satisfaction

We conducted our analysis next for the relationship satisfaction scale. Table 3 shows the comparisons of the different models for relationship satisfaction. Moving from the noninvariance model to the strong measurement-equivalent model resulted in a significant drop in model fit. All of the items showed high loadings on the latent factor, with the exception of the reverse-scored item ("I do not feel that my current relationship is successful"), and met the criteria for weak measurement invariance (i.e., equal factor loadings across cultural groups; see Table 4 for items). Assumptions for strong measurement invariance were violated for two of the five items in the Chinese sample and had to be relaxed in the final model

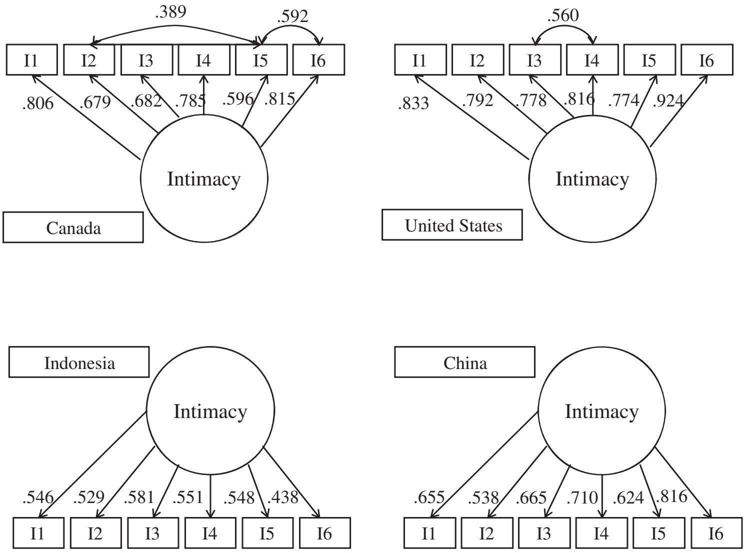

Figure 1. Final model of intimacy for each culture, showing the fully standardized model parameters.

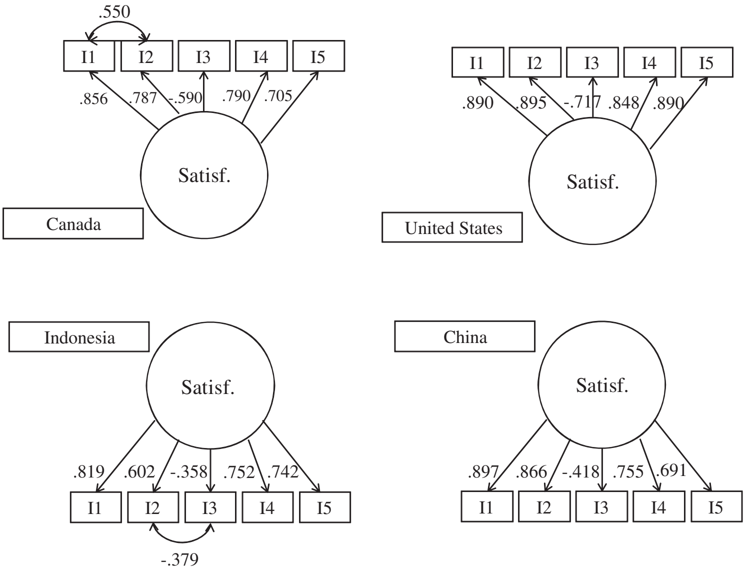

Figure 2. Final model of relationship satisfaction for each culture, showing the fully standardized model parameters.

("I do not feel that my current relationship is successful" and "My relationship with my partner is very rewarding"). When the assumption of equivalence was relaxed in the model for these two items for the Chinese sample, the resulting partial-invariance model had acceptable fit to the data, $\chi^2(40) = 63.60$, $p = .010$, CFI = .956, RMSEA = .077, SRMR = .119, and did not result in a decline in model fit compared to the baseline model. Figure 2 shows the final model for the full satisfaction scale with the fully standardized factor loadings. In constructing a shorter scale, when these two

Table 5. *Latent factor means, reliabilities (Raykov's rho), and tests of mean differences between samples using full scales*

Scales	Latent mean	Raykov's ρ	χ^2	$\delta\ \chi^2$	*p*
Intimacy					
American sample	.00	.91	84.02	—	—
Canadian sample	.44	.84	87.74	3.72	>.05
Indonesian sample	.13	.69	84.54	0.52	>.05
Chinese sample	.46	.82	87.92	3.9	~.05
Satisfaction					
American sample	.00	.93	63.60	—	—
Canadian sample	.60	.84	70.70	7.10	<.05
Indonesian sample	.52	.69	67.92	4.32	<.05
Chinese sample	.98	.80	78.63	15.03	<.05
Trust					
American sample	.00	.92	63.44	—	—
Canadian sample	.72	.86	73.18	9.74	<.05
Indonesian sample	.34	.76	65.47	2.03	>.05
Chinese sample	.90	.86	77.95	14.51	<.05
Commitment					
American sample	.00	.85	19.28	—	—
Canadian sample	.53	.64	25.12	5.84	<.05
Indonesian sample	−.29	.45	20.18	0.90	>.05
Chinese sample	.28	.57	20.79	1.51	>.05

items were eliminated, the remaining items met the criteria for strong measurement equivalence. The fit of this shorter scale was also acceptable, $\chi^2(12) = 21.45$, $p = .044$, CFI = .969, RMSEA = .089, SRMR = .100. Thus, for the satisfaction scale, although assumptions for weak measurement equivalence were met for all items, assumptions for strong invariance were not met for two items that were found to violate assumptions in the Chinese sample.

We examined mean differences on relationship satisfaction between the four cultural groups next, once again using the U.S. sample as the comparison group (Table 5). In these analyses, constraining the mean of any of the other three samples resulted in a significant drop in the chi-square value, indicating that all three samples had higher satisfaction scores than the U.S. sample. Finally, we tested whether the variance of relationship satisfaction scores was equivalent across the samples. When we constrained the variance of the latent satisfaction factor to be equivalent across the groups, the decrease in model fit was significant, indicating that the variance of satisfaction scores also differs across the groups.

Trust

We next conducted our analysis for the trust scale (see Table 3 for model comparisons). Moving from the noninvariance model to the strong measurement-equivalent model did not result in a significant change in model fit. All of the items showed high loadings on the latent factor and met the assumption for weak measurement invariance (i.e., equal factor loadings across cultural groups; see Table 4 for items). Two of the five items, however, did not meet assumptions for strong measurement equivalence. One of the items ("I feel that my partner can be counted on to help me") did not meet the assumption of strong invariance in the Chinese sample, and another item ("My partner does not hesitate to make significant sacrifices to

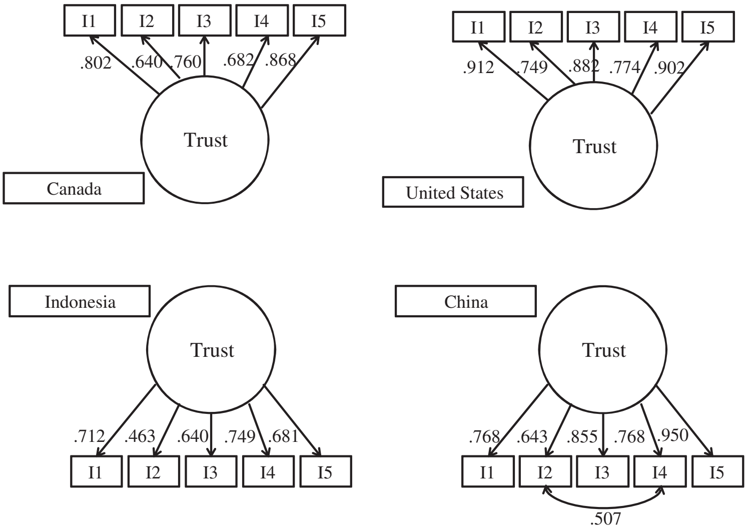

Figure 3. Final model of relationship trust for each culture, showing the fully standardized model parameters.

strengthen our relationship") did not meet the assumption in the Indonesian sample. When the assumption of equivalence was relaxed in the model for these two items, the resulting partial-invariance model had acceptable fit to the data, $\chi^2(41) = 63.44$, $p = .014$, CFI = .966, RMSEA = .074, SRMR = .125, and fit the data significantly better than the strong-invariance model. Figure 3 shows the final model for the full trust scale with the fully standardized factor loadings. In constructing a shorter scale, when these two items were eliminated, the remaining items met the criteria for strong measurement equivalence. The fit of this shorter scale was also acceptable, $\chi^2(12) = 10.45$, $p = .577$, CFI = 1.00, RMSEA = .000, SRMR = .091. Thus, for the trust scale, although assumptions for weak measurement equivalence were met for all items, two items did not meet strong invariance assumptions.

Next, we examined mean cultural differences in trust scores, using the U.S. sample as our comparison group (Table 5). Constraining the Canadian and the Chinese sample means to be equivalent to the mean of the U.S. sample resulted in significant drops in model fit. However, constraint on the mean of the Indonesian sample did not result in significant changes in model fit. Thus, the Canadian and Chinese samples had significantly higher trust scores than the U.S. sample, whereas the Indonesian sample did not. Finally, we tested whether the variance of trust scores was equivalent across the samples. When we constrained the variance of the latent trust factor to be equal across the groups, the decrease in model fit was significant, indicating that the variance of trust scores differs across the groups.

Commitment

We conducted our analysis next for the relationship commitment scale (see Table 3 for model comparisons). Moving from the non-invariance model to the strong measurement-equivalent model did not result in a significant change in model fit. All of the items showed high loadings on the latent factor, with the exception of the reverse-scored item ("I do not feel any moral duty or obligation to continue my relationship") in all samples but the U.S. sample; however, the factor loadings met the assumption for weak measurement invariance (i.e., equal factor loadings across cultural groups; see Table 4 for scale items). Assumptions for strong measurement invariance were also met for all three items. The fit of the

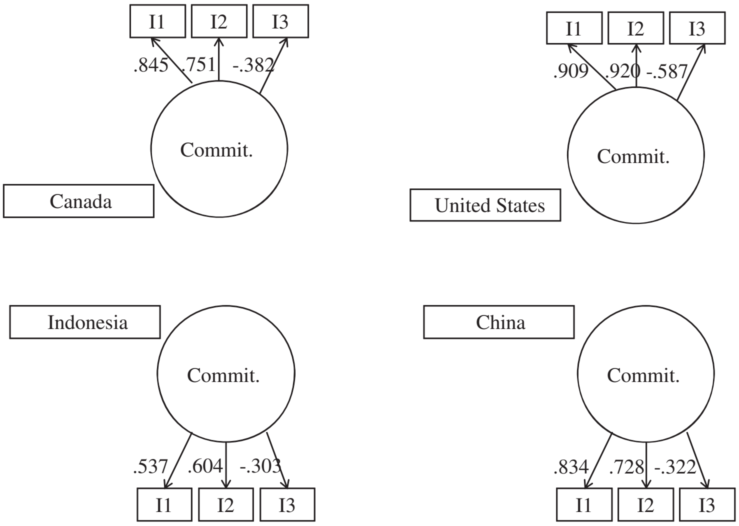

Figure 4. Final model of relationship commitment for each culture, showing the fully standardized model parameters.

strong measurement invariant model had acceptable fit to the data, $\chi^2(12) = 19.28$, $p = .082$, CFI = .918, RMSEA = .078, SRMR = .086. Figure 4 shows the final model for the commitment scale with the fully standardized factor loadings.

Next, we examined mean differences on commitment, using the U.S. sample as the comparison group (Table 5). Constraining the mean of the Canadian sample to be equal to the U.S. sample resulted in a significant drop in model fit. However, constraint of the Chinese and the Indonesian means did not result in significant changes in model fit. Thus, the Canadian sample reported significantly higher levels of commitment than the Indonesian and the U.S. sample. Finally, we tested whether the variance of commitment scores was equal across the samples. When we constrained the variance of the latent commitment factor to be equal across the groups, the decrease in model fit was significant, indicating that the variance of commitment scores differs across the groups.

Relations between constructs

As a final test of the scales, we wanted to test again whether the variance of the latent constructs was equal across the samples (referred to as latent factor variance invariance), as well as whether the constructs correlated with one another the same way across samples (referred to as latent factor covariance invariance). To test these, we created a model in which we included all four constructs (using the partial-invariance models). The baseline model did not include restrictions in the latent constructs across the samples (see Table 6 for latent construct correlations; see bottom of Table 3 for model comparisons). We then compared the baseline model with a model in which we constrained the variances of the latent factors across the samples, and to a model in which we constrained both the variances of the latent factors and the correlations between them to be equal across samples. As expected based on the prior tests of the individual scales, constraining the variances of the latent factors representing the constructs across the samples resulted in a significant decrease in model fit. This test once again indicated that the variances of the constructs are not equivalent across the samples. Further adding a constraint to hold the covariance of the constructs with one another equivalent across the samples also resulted in significant drop in model fit, compared to both the baseline and the invariant latent factor variance model. The results of this

Table 6. *Correlations between the latent factors with 95% confidence intervals using full scales*

Sample	1	2	3	4
American sample				
1. Intimacy	1.00	[0.69, 1.00]	[0.44, 0.86]	[0.38, 0.79]
2. Satisfaction	.84	1.00	[0.79, 0.95]	[0.53, 0.88]
3. Trust	.65	.87	1.00	[0.47, 0.85]
4. Commitment	.59	.71	.66	1.00
Canadian sample				
1. Intimacy	1.00	[0.36, 0.92]	[0.30, 0.92]	[0.03, 0.53]
2. Satisfaction	.64	1.00	[0.77, 0.98]	[0.34, 0.78]
3. Trust	.61	.88	1.00	[.20, .76]
4. Commitment	.28	.56	.48	1.00
Indonesian sample				
1. Intimacy	1.00	[0.71, 1.00]	[0.60, 0.97]	[0.51, 1.03]
2. Satisfaction	.86	1.00	[0.63, 0.99]	[0.71, 1.17]
3. Trust	.79	.81	1.00	[0.61, 1.26]
4. Commitment	.77	.94	.94	1.00
Chinese sample				
1. Intimacy	1.00	[0.58, 1.02]	[0.49, 0.96]	[0.42, 0.98]
2. Satisfaction	.80	1.00	[0.84, 0.98]	[0.41, 0.98]
3. Trust	.73	.91	1.00	[0.47, 0.99]
4. Commitment	.70	.69	.73	1.00

Note. Values below the diagonal represent the correlations between the latent constructs; values above the diagonal represent the 95% confidence intervals of the correlation estimates.

test indicate that the correlations among these constructs differ across samples.

Discussion

In this study, we examined the measurement equivalence of different scales of relationship quality to test whether these scales function the same way across four different countries. In addition, we tested differences in means and variances for each construct and whether the constructs relate to one another the same way across the four groups. We also attempted to establish scales that meet assumptions for strong measurement invariance for use when the goal of research is primarily to conduct tests of mean differences across groups. Our analyses indicated that all four scales tested met the criteria for weak measurement equivalence, in which factor loadings of the items on the latent construct are held to be equivalent across cultural groups. However, with the exception of the commitment scale, assumptions for strong measurement equivalence (i.e., equal intercepts) were violated for at least one item in each scale. For these items, it is possible to relax the requirement of equal intercepts to create partially equivalent scales. If these problematic items are eliminated from the scales, the remaining items meet criteria for strong invariance.

Cross-cultural researchers currently rely on back-translation, which is necessary but it is not a sufficient procedure for ensuring that a scale remains psychometrically valid (Borsboom, 2006; Byrne & Campbell, 1999). In addition to ensuring proper translation, the translated scale must also be thoroughly examined before it can be used to make conclusions about group differences across cultures. More specifically, researchers need to examine the measurement equivalence of the original and translated scales in order to conclude that the items function the same way across cultures (Borsboom, 2006; Byrne & Campbell, 1999; French & Finch, 2006;

Kline, 2011; Little, 1997; Meredith, 1993). Cross-cultural differences simply cannot be properly interpreted without establishing that the scale is equivalent. Unfortunately, equivalence is rarely investigated explicitly. We believe that this issue should be given more attention by those who are interested in cross-cultural research and that whenever possible, scales should be explicitly examined for measurement equivalence. The widespread availability of user-friendly statistical software for conducting multiple-group confirmatory factor analysis provides researchers with the means to conduct this analysis and should lead to examinations of measurement equivalence on a regular basis.

In this study, we conducted measurement equivalence analyses of four scales that tap into various aspects of relationship quality (i.e., intimacy, trust, relationship satisfaction, and commitment) that have often been the focus of research on relationships. Our goal was to provide researchers with a set of scales that could be used across cultures with some degree of confidence regarding their measurement properties. The results showed that all of the scales met the criteria for weak measurement equivalence. The scale for commitment also met assumptions for strong measurement equivalence, but these assumptions had to be relaxed for one or two items in the remaining three scales, leading to partial equivalence. Given that all of the scales met conditions for weak invariance, these scales can be used in cross-cultural research in the future for most research questions. Weak invariance is a necessary condition for cross-cultural comparisons because it represents construct validity (French & Finch, 2006). Given the establishment of weak invariance, as scores on the construct change in one culture, the scores on these scales change the same way in the other cultures. Thus, the scales can be used to examine relations between the constructs and other variables of interest to relationship researchers. We hope that when researchers are restricted by small sample sizes that do not allow conducting multiple-group confirmatory analyses, they will try to use scales that have been tested for and show measurement equivalence in prior research, such as those included in this study.

It is important to note that because three of the scales did not meet assumptions of strong measurement invariance (satisfaction, intimacy, and trust), they may not be appropriate for use when the goal of research is to compare mean scores across cultures. However, our results suggest that the commitment scale tested here may be appropriate for such comparisons (although reliabilities for this scale were low in the Chinese and Indonesian samples). Strong invariance is necessary when cultural groups are compared on mean levels of a construct, such as when researchers are interested in differences in marital satisfaction across cultures (Borsboom, 2006; Byrne & Campbell, 1999; French & Finch, 2006; Kline, 2011; Meredith, 1993; Slof-Op 't Landt et al., 2009). To address this problem with the three scales that did not meet the requirements for strong measurement invariance, we attempted to remove the problematic items from the scales to construct shorter scales that did demonstrate strong measurement invariance and thus could potentially be used for mean-level comparisons. However, elimination of items from a scale may result in problems with both reliability and validity (if the remaining items do not capture the essence of the construct). We compared these shorter scales with the longer versions on mean scores, standard deviations, scale reliability, and correlations with self-esteem to provide tests of discriminant validity (see Table 1). Although mean scores, standard deviations, and correlations with self-esteem were similar for the long and short scales for all of the constructs, problems with reliability emerged for the intimacy and trust scales in the Indonesian sample. For the satisfaction scale, the reliabilities remained very similar to the values of the longer scales. In light of the drop in reliabilities, researchers should be cautious if they wish to use these shorter scales in future research. Future work may be needed to establish the reliability and validity of these shorter scales for use in cross-cultural research when the goal is to compare construct means across different cultural groups. However, the commitment scale met strong measurement invariance, making it appropriate for use when testing mean-level differences, although it is important to point out that in the Chinese and Indonesian samples, reliability of this scale was quite low. We suggest that researchers examine the reliability of this scale in their own samples first and use the scale to test mean-level differences only if reliability is adequate in their samples.

Why did some items fail to meet assumptions of strong equivalence? There are several reasons why an item may not be invariant across cultural groups (Byrne & Campbell, 1999). Sometimes, despite the best efforts of researchers, translation errors may be at play. However, even when items are properly translated, the behaviors tapped by the item may not be appropriate in a particular cultural context. This issue relates to what has been referred to as conceptual equivalence in constructs. When a construct is conceptually equivalent across cultures, the construct means the same thing and is represented similarly (e.g., through attitudes and behaviors) across cultures (Hui & Triandis, 1985; Singh, 1995). Ideally, researchers should pay attention to such issues when constructing scales and make sure that the attitudes and behaviors represented by the items are equally appropriate in different cultural settings (Byrne et al., 2009; Singh, 1995). For example, in this study, the intimacy item of sharing deeply personal information about oneself may have been problematic in the Chinese sample because it may not be as appropriate to share this type of information with one's partner in Chinese culture as it is in the other three cultures tested here. In fact, there is existing evidence that self-disclosure, even to intimate friends, is less appropriate in Chinese society compared to other nations (Chen, 1995). Thus, this item in particular may have been one that is not equally appropriate across the cultures and has led to lack of measurement invariance. However, when lack of invariance is found, it is often difficult to be certain of what the exact cause may be.

After we examined the scales for measurement equivalence, we also tested whether there are differences across the cultural groups in the means and the variances of the constructs. We did not find evidence of mean differences in intimacy, but on the other scales there was evidence of some mean differences. For all of the scales, the variances of the construct were not equal across the samples. Although the focus in this article was on showing what types of tests can be conducted using measurement-equivalent scales rather than testing specific hypotheses regarding differences across cultures, it is important to note that when such differences between cultures emerge using measurement-equivalent scales, researchers can begin to seek explanations for the differences. In our study, the U.S. sample scored lower on satisfaction and trust than the other groups, which in this case may be due to the fact that this sample was recruited through the Internet, rather than through in-person recruitment. Thus, it is possible that the characteristics of the samples influenced the mean scores on the constructs and these are important issues to pay attention to when researchers are interested in making mean-level comparisons across cultures and draw conclusions from any emerging differences.

In addition to testing differences in means and variances, we also tested whether the constructs relate to one another the same way across the cultures. Importantly, all of the scales correlated positively with one another, demonstrating convergent validity between the scales across cultures. However, correlations between the scales differed across the samples. In particular, it is noteworthy that in the Chinese and Indonesian samples some of the correlations between the constructs were quite high. These high correlations between the constructs may indicate that these indicators of relationship quality are less differentiated in more collectivistic cultures than in more individualistic cultures. Such questions of construct differentiation are also important to address in cross-cultural research and relate to the issue of functional equivalence. Functional equivalence is present when constructs relate to other variables across cultures the same way, such that they have similar consequences and are influenced by the same causal forces (Hui & Triandis, 1985; Singh, 1995). In other words, the constructs have the same function across cultures (Singh, 1995). In our experience, these are the types of questions that relationship researchers are interested in most of the time, as this is when associations between variables are compared across cultures. For these types of research questions, the scales used in this study are appropriate given that they all met requirements for weak measurement invariance.

It is worth noting that in our study, we tested invariance only across four cultural groups, which by no means provides an exhaustive test of all possible cultural groups. When the scales tested in this study are used in other cultures, researchers should try to test them for invariance if possible. However, when testing invariance is not feasible, we still recommend the use of scales that have at least been examined for invariance and have demonstrated good measurement properties in both individualistic and collectivistic samples. We selected two groups (United States and Canada) that can be considered to be high on individualism and two that are higher on collectivism (Hofstede, 1980) with this aim in mind.

In sum, we believe that the growing focus on cross-cultural comparisons necessitates simultaneous examinations of measurement equivalence. Findings of cultural differences cannot be clearly interpreted without considering the invariance of the scales used across cultural groups. Relationship scholars should begin to address measurement equivalence in studies with multiple cultural groups. When it is not possible to do so, it is important to rely on scales that have already been examined for invariance and have been demonstrated to hold up well in such analyses. In this study, we provide scales to assess several aspects of relationship quality that relationships researchers can use in their own studies. The use of such scales can increase our confidence that reported results reflect differences that are most likely a result of actual differences between the cultural samples, rather than problems with the measurement properties of the scales used.

References

Aron, A., Aron, E. N., & Smollan, D. (1992). Inclusion of other in the self scale and the structure of interpersonal closeness. *Journal of Personality and Social Psychology, 63*, 596–612. doi: 10.1037/0022-3514.63.4.596

Baumeister, R. F., & Leary, M. R. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin, 117*, 497–529. doi: 10.1037/0033-2909.117.3.497

Bentler, P. M., & Chou, C. (1987). Practical issues in structural modeling. *Sociological Methods and Research, 16*, 78–117.

Borsboom, D. (2006). When does measurement invariance matter? *Medical Care, 44*, S176–S181. doi: 10.1097/01.mlr.0000245143.08679.cc

Bradbury, T. N., Fincham, F. D., & Beach, S. R. (2000). Research on the nature and determinants of marital satisfaction: A decade in review. *Journal of Marriage and Family, 62*, 964–980.

Byrne, B. M., & Campbell, T. L. (1999). Cross-cultural comparisons and the presumption of equivalent measurement and theoretical structure—A look beneath the surface. *Journal of Cross-Cultural Psychology, 30*, 555–574.

Byrne, B. M., Oakland, T., Leong, F. T. L., van de Vijver, F. J. R., Hambleton, R. K., Cheung, F. M., & Bartram, D. (2009). A critical analysis of cross-cultural research and testing practices: Implications for improved education and training in psychology. *Training and Education in Professional Psychology, 3*, 94–105.

Chen, G. (1995). Differences in self-disclosure patterns among Americans versus Chinese: A comparative study. *Journal of Cross-Cultural Psychology, 26*, 84–91.

Dion, K. K., & Dion, K. L. (1993). Individualistic and collectivistic perspectives on gender and the cultural context of love and intimacy. *Journal of Social Issues, 49*, 53–69.

Finkel, E. J., Rusbult, C. E., Kumashiro, M., & Hannon, P. A. (2002). Dealing with betrayal in close relationships: Does commitment promote forgiveness? *Journal of Personality and Social Psychology, 82*, 956–974. doi: 10.1037/0022-3514.82.6.956

French, B. F., & Finch, W. H. (2006). Confirmatory factor analytic procedures for the determination of measurement invariance. *Structural Equation Modeling, 13*, 378–402.

Gere, J., & MacDonald, G. (2010). An update of the empirical case for the need to belong. *Journal of Individual Psychology, 66*, 93–115.

Goodwin, R., & Findlay, C. (1997). "We were just fated together." Chinese love and the concept of yuan in England and Hong Kong. *Personal Relationships, 4*, 85–92.

Heller, D., Watson, D., & Ilies, R. (2004). The role of person versus situation in life satisfaction: A critical examination. *Psychological Bulletin, 130*, 574–600. doi: 10.1037/0033-2909.130.4.574

Hofstede, G. (1980). *Culture's consequences: International differences in work-related values*. Beverly Hills, CA: Sage.

Hui, C. H., & Triandis, H. C. (1985). Measurement in cross-cultural psychology: A review and comparison of strategies. *Journal of Cross-Cultural Psychology, 16*, 131–152.

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: Guilford.

Lalonde, R. N., Hynie, M., Pannu, M., & Tatla, S. (2004). The role of culture in interpersonal relationships: Do second generation South Asian Canadians want a traditional partner? *Journal of Cross-Cultural Psychology, 35*, 503–524.

Larzelere, R. E., & Huston, T. L. (1980). The Dyadic Trust Scale: Toward understanding interpersonal trust in close relationships. *Journal of Marriage and Family, 42*, 595–604.

Laurenceau, J., Barrett, L. F., & Rovine, M. J. (2005). The interpersonal process model of intimacy in marriage: A daily-diary and multilevel modeling approach. *Journal of Family Psychology, 19*, 314–323.

Laurenceau, J., & Kleinman, B. M. (2006). *Intimacy in personal relationships*. New York, NY: Cambridge University Press.

Le, B., & Agnew, C. R. (2003). Commitment and its theorized determinants: A meta-analysis of the investment model. *Personal Relationships, 10*, 37–57. doi: 10.1111/1475-6811.00035

Levine, R., Sato, S., Hashimoto, T., & Verma, J. (1995). Love and marriage in eleven cultures. *Journal of Cross-Cultural Psychology, 26*, 554–571.

Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research, 32*(1), 53–76. doi: 10.1207/s15327906mbr3201_3

MacDonald, G., & Jessica, M. (2006). Family approval as a constraint in dependency regulation: Evidence from Australia and Indonesia. *Personal Relationships, 13*, 183–194.

Marshall, T. C. (2008). Cultural differences in intimacy: The influence of gender-role ideology and individualism-collectivism. *Journal of Social and Personal Relationships, 25*, 143–168.

McAdams, D. P. (1985). Motivation and friendship. In S. Duck & D. Perlman (Eds.), Understanding personal relationships: An interdisciplinary approach (pp. 85–105). London, England: Sage.

McGrath, R. E. (2009). On prototypes and paradigm shifts. *Measurement, 7*, 27–29.

Meredith, W. (1993). Measurement invariance, factor-analysis and factorial invariance. *Psychometrika, 58*, 525–543.

Murray, S. L., & Holmes, J. G. (2009). The architecture of interdependent minds: A motivation-management theory of mutual responsiveness. *Psychological Review, 116*, 908–928. doi: 10.1037/a0017015

Murray, S. L., Holmes, J. G., & Griffin, D. W. (1996a). The benefits of positive illusions: Idealization and the construction of satisfaction in close relationships. *Journal of Personality and Social Psychology, 70*, 79–98.

Murray, S. L., Holmes, J. G., & Griffin, D. W. (1996b). The self-fulfilling nature of positive illusions in romantic relationships: Love is not blind, but prescient. *Journal of Personality and Social Psychology, 71*, 1155–1180.

Muthén, L., & Muthén, B. (2007). *MPlus 5*. Los Angeles, CA: Muthén & Muthén.

Oyserman, D., Coon, H. M., & Kemmelmeier, M. (2002). Rethinking individualism and collectivism: Evaluation of theoretical assumptions and meta-analyses. *Psychological Bulletin, 128*, 3–72.

Prager, K. J., & Buhrmester, D. (1998). Intimacy and need fulfillment in couple relationships. *Journal of Social and Personal Relationships, 15*, 435–469.

Rempel, J. K., Holmes, J. G., & Zanna, M. P. (1985). Trust in close relationships. *Journal of Personality and Social Psychology, 49*, 95–112.

Rusbult, C. E. (1980). Commitment and satisfaction in romantic associations: A test of the investment model. *Journal of Experimental Social Psychology, 16*, 172–186. doi: 10.1016/0022-1031(80)90007-4

Rusbult, C. E. (1983). A longitudinal test of the investment model: The development (and deterioration) of satisfaction and commitment in heterosexual involvements. *Journal of Personality and Social Psychology, 45*, 101–117. doi: 10.1037/0022-3514.45.1.101

Singh, J. (1995). Measurement issues in cross-national research. *Journal of International Business Studies, 26*, 597–619.

Slof-Op 't Landt, M. C. T., van Furth, E. F., Rebollo-Mesa, I., Bartels, M., van Beijsterveldt, C. E. M., Slagboom, P. E., & Dolan, C. V. (2009). Sex differences in sum scores may be hard to interpret the importance of measurement invariance. *Assessment, 16*, 415–423.

Snyder, D. K., Castellani, A. M., & Whisman, M. A. (2006). Current status and future directions in couple therapy. *Annual Review of Psychology, 57*, 317–344.

Sternberg, R. J. (1997). Construct validation of a triangular love scale. *European Journal of Social Psychology, 27*, 313–335.

Tsapelas, I., Aron, A., & Orbuch, T. (2009). Marital boredom now predicts less satisfaction 9 years later. *Psychological Science, 20*, 543–545. doi: 10.1111/j.1467-9280.2009.02332.x

Van Lange, P. A. M., Rusbult, C. E., Drigotas, S. M., Arriaga, X. B., Witcher, B. S., & Cox, C. L. (1997). Willingness to sacrifice in close relationships. *Journal of Personality and Social Psychology, 72*, 1373–1395. doi: 10.1037/0022-3514.72.6.1373

Whitton, S. W., & Kuryluk, A. D. (2012). Relationship satisfaction and depressive symptoms in emerging adults: Cross-sectional associations and moderating effects of relationship characteristics. *Journal of Family Psychology, 26*, 226–235.