Four-dimensional hierarchical structure of love constructs in a cross-cultural perspective

Karandashev, Victor; Evans, Nicholas D.; Neto, Félix; Zarubko, Elena; Artemeva, Veronika; Fallah, Sadeq; Cassepp#Borges, Vicente; Nazaré Oliveira, Eliany; Akyol, Havvanur; Dincer, Duygu

INTERNATIONAL ADAPTATION OF MEASUREMENT INSTRUMENTS   **Open Access**

Four-dimensional hierarchical structure of love constructs in a cross-cultural perspective

Victor Karandashev[1*], Nicholas D. Evans[2], Félix Neto[3], Elena Zarubko[4], Veronika Artemeva[5], Sadeq Fallah[6], Vicente Cassepp-Borges[7], Eliany Nazaré Oliveira[8], Havvanur Akyol[9] and Duygu Dincer[10]

Abstract

This article reports new methodology for cross-cultural exploration of psychometric properties of a four-dimensional hierarchical love scale. We collected data from 2831 participants from nine regional locations from six countries and assessed their responses to the love scale as well as several other love feelings. We applied a new methodological approach using recently advanced statistical methods to the comparison of forty love attitudes underscoring four distinct latent attitudes associated with love to another person in romantic relationships across these samples. The results demonstrate the importance of measurement invariance tests for cross-cultural comparison of scores on love scales. To properly assess measurement invariance, we suggest five statistical procedures, which we investigated in this study: (1) making corrections for acquiescence and extreme response biases; (2) taking into consideration cultural uniqueness in how participants respond to the measures, which may contribute to poor model fit; (3) accounting for such cultural uniqueness to make cross-cultural comparisons more valid; (4) removing items, which substantially contribute to poor model fit; and (5) shortening the subscales when scoring and analyzing the data. The results of the studies propose two shortened versions (33 and 30 items) of the love scale as two cross-culturally valid and invariant alternatives to the original 40-item scale.

Keywords: Love constructs, Derived etic approach, Methodology of analysis, Psychometrics of love scale, Measurement invariance across samples

The aim of the current study was to apply a new methodological approach for a cross-cultural comparison of psychometric structures of love attitudes as multidimensional constructs. We assessed the attitudes associated with romantic love using data collected from several cultural samples (from the USA, Portugal, Russia, Brazil, Iran, and Turkey). The results of this innovative methodological analysis can be useful for testing cross-cultural invariance of scales measuring emotions and attitudes in relationships using data obtained from several cultural groups.

Theoretical background
Methodology of multidimensional and cross-cultural love research
Multidimensionality of love construct
An extensive examination of multiple love studies, which have used psychometric scales since the 1960s (see for review Karandashev & Evans, 2019), showed that love is a multifaceted construct consisting of many attitudes and emotions. Forty of such love constructs are especially prominent in painting a comprehensive picture of love. While some were left outside of contemporary love research, we believe that they deserve their spot in the modern comprehensive theory of love. The theory defines those characteristics of love — basic love attitudes — as distinguishable, yet interdependent constructs. From a theoretical review and analysis, those 40

*Correspondence: vk001@aquinas.edu

[1] Aquinas College, Grand Rapids, USA
Full list of author information is available at the end of the article

constructs characterizing various aspects of individual love attitudes meaningfully fit into four dimensions of love: compassion, affection, closeness, and commitment.

Compassion is associated with care, concern, empathy, acceptance, tolerance, protection, giving advice, consolation, support, and nonjudgment. *Affection* is associated with admiration, respect, personality attraction, tenderness, elation, appreciation, compliment, physical embrace, considerate, and physical attraction. *Closeness* is associated with attachment, openness, affiliation, compatibility, feeling understood, feeling accepted, reliance, trust, emotional comfort, and seeking help. *Commitment* is associated with devotion, desire for relationship, long-term orientation, exclusiveness, stability, forgiveness, coping, investment, sacrifice, and cooperation. These four groups of love attitudes (first-order constructs) can be classified in the two second-order constructs of love: (a) *the attitudes toward the partner — compassion* and *affection*, and (b) *the attitudes of the relationship — closeness* and *commitment*.

We are aware that this list of basic love attitudes and their relations is a matter of personal and cultural interpretations. Love is overall quite subjective. Nevertheless, these 40 constructs have been most widely acknowledged by researchers and studied in love scholarship over the last 60 years. Each item of the scale describes its corresponding construct as a statement expressing personal attitude toward another person and relationship with them.

Exploratory and confirmatory factor analyses (Karandashev & Evans, 2019) supported the four-dimensional structure of those items in two American samples (Midwest and Southeast) and one British sample. The 40 items describing basic love attitudes distinctively loaded on their corresponding four first-order dimensions, and the item-total correlations were distinctively higher on those dimensions. The results of the psychometric analysis replicated in two studies show stability of the factor structure.

However, due to cross-cultural diversity of love conceptions (see, for review, Karandashev, 2017, 2019), such a four-dimensional structure of the 40 basic love constructs may not be culturally universal. Even though the constructs of compassion, affection, closeness, and commitment are present in many cultures (Karandashev, 2017, 2019), their specific emotional content and associations may have different connotations. Therefore, the purpose of this study was to test the cross-cultural universality of the four-dimensional hierarchical model of love attitudes originally developed by Karandashev and Evans (2019). We expected that the latent love constructs and their corresponding items/attitudes should show cross-cultural similarity, while others show cultural specificity.

Challenges of psychometrics in love research

Previous studies have revealed (see for review Karandashev, 2019; Karandashev & Evans, 2019) that many traditional psychometric methods are not very suitable for analysis of love scales. Therefore, psychometric structures and data are difficult to interpret, especially in cross-cultural studies. The interpretations often appear confusing and not quite adequate, and comparisons between cultural samples are barely meaningful.

Among those problems are the evident tendencies to extreme responses and acquiescence as well as observed high inter-item and item-total correlations between both items within and across latent constructs. For example, the scores of dimensions of the original 40-item scale (Karandashev & Evans, 2019) were intercorrelated, which is also a common occurrence among most other love constructs and dimensions (e.g., Fehr, 1994; Graham, 2011; Masuda, 2003; see for review, Karandashev, 2019, Karandashev & Evans, 2019). This may indicate issues related to extreme responding and acquiescence (likely both a result of halo effects), in addition to interdependence between the constructs. Although these response sets are understandable for such highly passionate attitudes as love, the self-report ratings of items in love scales are skewed to the high-level range with high density of distribution. Commonly, there is no normal distribution in data sets (see for review and explanation, Karandashev, 2019; Karandashev & Evans, 2019).

Finally, traditional methods of psychometrics (e.g., exploratory, confirmatory factor analysis) encounter problems when identifying distinct factors. There appears to be a great deal of overlap between love attitude constructs (e.g., Fehr, 1994; Graham, 2011; Masuda, 2003; see for review and detail, Karandashev, 2019, 2021). Therefore, new research methodology and statistical analysis are needed to explore the multidimensionality of love.

Measurement invariance in cross-cultural love research

Love researchers encounter another set of methodological challenges in cross-cultural studies. Self-report survey methods are often analyzed using several statistical methods and criteria to test measurement invariance across cultural samples of interest (e.g., Davidov et al., 2014; Johnson, 1998; Milfont & Fischer, 2010; Van De Schoot et al., 2015; Van de Vijver & Leung, 2011). Cross-cultural love studies have not applied this new methodology in their data analysis so far.

Traditionally, researchers developing and validating love scales in other cultures present its psychometrics — factor structure, reliability, and validity in their cultural samples. Moreover, they often imply that their

measures are invariant across samples. However, these cross-cultural love studies usually do not test the cross-cultural measurement invariance of their data — with the assumption that basic psychometric tests of validity and reliability equate to measurement invariance. This assumption, however, is not adequate. The psychometrics of a scale depend not only on the scale's qualities but also on the sample's characteristics and can, therefore, differ widely across samples. In other words, as participants from a given sample may differ on important demographic characteristics from participants from another sample, the samples used to test the validity and reliability of a measure can affect both the validity and reliability of the measure and its invariance across groups. Therefore, love researchers should ensure that psychometric properties of scales are equivalent, stable, and invariant across different groups and/or measurement times (see for review and guide, Davidov et al., 2014; Johnson, 1998; Vandenberg & Lance, 2000; Van De Schoot et al., 2015; Van de Vijver & Leung, 2011).

It is important to test and verify measurement invariance when group characteristics (e.g., cultural membership) may influence how participants respond to the measure. This helps avoid measurement bias and allows researchers to compare the scores of latent constructs across different groups. Otherwise, comparison of means is not adequate. However, in many cross-cultural studies (see for review Boer et al., 2018; Fischer & Karl, 2019; Van De Schoot et al., 2015), the means for latent variables are compared between groups without sufficient psychometric basis. Unfortunately, researchers infrequently run proper psychometric analyses before comparing variables between groups when they use the scale in the same language and within the same culture.

Typological, demographic, or cultural characteristics of participants can contribute to potential measurement variance. Therefore, when researchers test measurement invariance, they rarely obtain the same parameters of measurement across cultural samples. For large data sets, poor fit is a standard finding (Byrne & Van de Vijver, 2017). These difficulties to meet the assumption of measurement invariance with psychometric scales across groups (see for review Van De Schoot et al., 2015) might be a reason why they often overlook this problem. When the violation of measurement invariance is not severe, the comparisons of data across samples may still be meaningful (Davidov et al., 2014).

Dealing with cross-cultural non-invariance in love scales
To verify measurement invariance, researchers—including cross-cultural researchers (e.g., Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000)—utilize the methods of multigroup confirmatory factor analysis

(Asparouhov & Muthén, 2014; Chen, 2008; Fischer & Karl, 2019; Hirschfeld & Von Brachel, 2014; Hui & Triandis, 1985). If the criteria of invariance are not met, researchers locate the sources of misfit, DIF, or other problems, and employ exploratory structural equation modeling (ESEM), the alignment method (CFA-derived technique), ordinal regression (item response theory) for differential item functioning (DIF), logistic regression, exploratory factor analysis (EFA), or principal component analysis (PCA) with Procrustes rotation with two or several groups (Fischer & Karl, 2019). When the source of measurement variance across samples is detected, researchers exclude variant items, constructs, or samples from analysis (Davidov et al., 2014; Fischer & Karl, 2019) to improve measurement equivalence and invariance.

Investigation of cross-cultural measurement invariance is important in the studies of emotions because people across cultures can ascribe different meanings to such emotions as anger, happiness, and love (see for review Karandashev, 2017, 2019, 2021). Although the core meanings of emotional constructs are substantially common, specific cultural contents of emotional constructs may differ (see for review Karandashev, 2021). Therefore, love can have universally basic components that are equivalent across cultural samples, while other components can be culturally specific.

Therefore, it is important that psychometric scales are both cross-culturally equivalent and invariant and culturally sensitive. Psychometricians should acknowledge that the scales can still be variant across cultures. Van de Vijver and Leung (1997) suggested using structural equation modeling and item response analysis for these purposes.

Aims and research questions of the studies
The aim of this study was to explore the structure of 40 basic constructs, which researchers have used to define the concept of love throughout the last 60 years. These constructs and corresponding items describe love attitudes of an individual toward another person within the context of various relationships (e.g., romantic, platonic, companionate, familial). Earlier studies (Karandashev & Evans, 2019) identified that these 40 basic dimensions are structured in four first-order and two second-order dimensions.

For this study, we expected that that this four-dimensional hierarchical structure of the basic dimensions is generally cross-culturally similar while still acknowledging cultural uniqueness. We sought to identify the degree of invariance in the structure of those 40 basic attitudes across cultural samples. We also investigated several key factors that affect invariance and consequently their

Table 1 Descriptive characteristics of each cultural sample

Sample	Gender *n*			Relationship status		Age		QLS factors *M (SD)*, α				Degree of "liking"	
	M	*W*	Tot.	Current	Previous	*M*	*SD*	Comp.	Affec.	Close.	Comm.	*M*	*SD*
Br.	66	241	307	246	61	25.56	7.13	6.17 (0.76), 0.87	6.30 (.87), 0.92	5.94 (1.19), 0.94	5.75 (1.12), 0.91	4.55	0.78
Ir.	117	236	353	244	109	23.54	4.29	4.96 (0.72), 0.85	5.11 (0.83), 0.91	4.87 (0.87), 0.89	4.85 (0.89), 0.89	N/A	N/A
Pt.	88	201	289	191	98	21.33	3.42	6.40 (0.68), 0.89	6.43 (0.82), 0.93	6.11 (0.99), 0.92	5.80 (1.04), 0.89	4.62	0.73
S.P.	168	177	345	222	123	21.27	3.99	5.89 (1.01), 0.91	5.89 (1.18), 0.95	5.75 (1.22), 0.94	5.60 (1.24), 0.92	4.42	0.78
Ty.	160	209	369	241	128	21.00	3.90	5.97 (0.92), 0.88	6.05 (1.07), 0.94	5.87 (1.14), 0.92	5.65 (1.17), 0.91	4.44	0.87
Hi.	78	201	279	155	124	19.97	2.77	6.26 (0.73), 0.89	6.21 (0.88), 0.93	5.97 (1.03), 0.93	5.70 (1.18), 0.92	4.45	0.84
MW.	41	97	138	81	57	19.09	1.54	6.30 (0.70), 0.89	6.20 (0.91), 0.93	6.00 (1.06), 0.94	5.81 (1.20), 0.94	4.51	0.70
SE.	160	317	477	273	204	19.99	2.73	6.25 (0.81), 0.91	6.17 (0.96), 0.93	5.93 (1.05), 0.92	5.74 (1.19), 0.93	N/A	N/A
Tur.	133	141	274	274	0	24.23	2.56	6.51 (0.36), 0.75	6.70 (0.34), 0.82	6.61 (0.39), 0.81	6.60 (0.40), 0.81	N/A	N/A
Tot.	1011	1820	2831	1927	904	21.83	4.42	6.05 (0.89), 0.91	6.09 (1.01), 0.94	5.87 (1.11), 0.93	5.71 (1.16), 0.92	4.51	0.80

Br. Brazil, *Ir.* Iran, *Pt.* Portugal, *S.P.* Saint Petersburg, Russia, *Ty.* Tyumen, Russia, *Hi.* Hawaii, USA, *MW.* Midwest, USA, *SE.* Southeast, USA, *Tur.* Turkey, *M* men, *W* women, *Tot.* total; current, participants currently in romantic relationships; previous, participants who have previously been in relationships, *Comp.* compassion, *Affec.* affection, *close.* closeness, *Comm.* commitment

structure. Ultimately, since this investigation was largely exploratory, we had four broader research questions.

- Research question 1: Does the cross-cultural adjustment of rating scores for acquiescence and extreme response biases increase invariance of the love scale?
- Research question 2: Does increasing the number of cultural samples increase variance of the scale? Do some samples increase variance more than others? If so, this would characterize the cultural similarities and differences in participants' understanding of the meaning of these basic constructs.
- Research question 3: Are some of the 40 dimensions associated with love cross-culturally invariant and compound into the four first-order factors-attitudes — affection, compassion, closeness, and commitment, and two second-order factors — the attitudes toward a partner (affection, compassion) and the attitudes regarding a relationship (closeness, commitment)? Are the remaining basic construct dimensions cross-culturally variant, reflecting their culturally specific meanings?
- Research question 4: Will elimination of variant items increase invariance of the scale, and, in turn, create more cross-culturally invariant shortened versions, allowing for cross-cultural comparison?

Instrument and scale development
Method
Participants and procedure
We collected data from 2831 student participants (1011 men, 1820 women; age: $M = 21.83$, $SD = 4.42$) from nine

different cultural regional locations across six different countries from around the world: Brazil, Iran, Portugal, Russia (in which two cultural samples were collected), Turkey, and the USA (in which three cultural samples were collected). As our factor analyses use maximum likelihood estimation and thus account for missing data, we included participants who may have not completed certain items of the scale. See Table 1 for descriptive statistics for each regional sample. All participants completed a series of questionnaires in which the measure of interest — the QLS (Karandashev & Evans, 2019) — was included. Upon completion of the survey, participants were granted credit for their respective psychology courses.

Measures
All participants completed the QLS (Karandashev & Evans, 2019), rating their attitudes toward a romantic partner— current or past — or acquaintance of the opposite gender. The scale consists of 40 items that are presented as statements, each of which describes a distinct love attitude. Karandashev and Evans' (2019) earlier work has demonstrated that these items assess four first-order dimensions of love feelings: compassion, affection, closeness, and commitment. Each factor includes 10 theoretically distinct feelings or attitudes, such as "I would console this person in times of need" (compassion), "I like to physically embrace this person" (affection), "I am comfortable asking this person for help" (closeness), and "I want to be in this relationship" (commitment). Participants read the following instructions:

> *Please rate your feelings toward your current romantic partner (if you are currently in a relation-*

ship), toward your most recent romantic partner (if you are currently not in a relationship but have been before), or toward your acquaintance of the opposite sex (if you have never been in a romantic relationship).

In the American samples, we used the original English version of QLS. In other countries, researchers translated and adapted the scale to their national languages following the recommended procedures of Hambleton and Zenisky (2010). Although we instructed participants who may have never been in a romantic relationship to rate their feelings toward an acquaintance, we included in the final data set the participants who were either currently in a romantic relationship or have been in one in the past. Therefore, the analyses included only participants rated their feelings toward their current romantic partner or past romantic partner on the 7-point rating scale (1 = "disagree strongly," 7 = "agree strongly"). See Table 1 for descriptive statistics of each factor for each regional sample as well as among the combined sample. See Table 1 of the Supplementary materials for means, standard deviations, and intercorrelations of the composite factors and factor items.

Analytic plan

We followed previously developed statistical procedures to test the cross-cultural psychometric properties of a self-report measure. We first opted to control for response sets during the analysis phase. Two primary response sets contribute to poor scale measurement: acquiescence and extreme responding (He & van de Vijver, 2012; Morren et al., 2012; van Herk et al., 2004). We created indices of both acquiescence and extreme responding using an approach used by Bachman and O'Malley (1984) as well as van Herk et al. (2004). Specifically, to calculate an index of acquiescence, we counted how many positive scores (i.e., values 6 and 7 on the 7-point scale) and negative scores (i.e., values 1 and 2) each participant had. The number of negative scores for each participant was then subtracted from the number of positive scores. This value was then divided by the number of items in the scale (40). This gave us an acquiescence index value from -1 (negative responding) to $+1$ (positive responding). To calculate an index of extreme responding, we counted how many times each participant responded using an anchor response along the 7-point scale (i.e., providing values of 1 or 7). This total value was then divided by the number of items (40) to give us an index value from 0 (no extreme responding) to 1 (complete extreme responding).

We then tested how well the scale measures what it is supposed to measure across diverse samples by testing

measurement invariance of the scale using a multigroup confirmatory factor analysis (MGCFA) procedure (Stein et al., 2006). Measurement invariance implies that the selected model of the scale is a good fitting model which generally does not vary across samples of interest. Since MGCFA only provides one measure of model fit, we tested which groups provide better or worse model fit. Therefore, individual group CFAs were implemented to estimate how well the scale applies to each cultural group. If a certain sample is shown to have substantially poorer model fit, we can then test the overall multigroup CFA model fit with that group omitted from analysis. For these purposes, we employed both the MGCFA and individual CFA approach in our analyses.

We measured absolute fit using the Yuan-Bentler corrected χ^2 test and three other measures: standardized root-mean-square residual (SRMR), the comparative fit index (CFI), and the root-mean-square error of approximation (RMSEA; Hu & Bentler, 1998, 1999). We compared each of the three indices of model fit to the model fit cutoffs (Hu & Bentler, 1999): SRMR ≤ .08, CFI ≥ 0.90), and RMSEA ≤ .06 or .08.

However, the use of strict cutoffs for fit indices has been a topic of debate (e.g., Kenny, 2012; McNeish et al., 2018). Hu and Bentler (1999) proposed the model fit cutoffs based on a simple model containing 15 indicators and three factors. At the same time, they suggested that CFA on multidimensional models needs to consider not only these golden rules but also a theoretical model to determine such cutoffs. The same requirements applicable for short and simple models are not appropriate to the lengthy and complex measures (Hopwood & Donnellan, 2010). For example, many studies in personality and sport psychology (see for review, Hopwood & Donnellan, 2010; Marsh et al., 2010; Perry et al., 2015) reviewed the suitability of using these cutoff values and showed that it is very difficult, sometimes impossible, to achieve acceptable fit for multidimensional measures. Moreover, as McNeish et al. (2018) noted, a RMSEA value of 0.06, usually suggesting good fit, can show poor fit and poor measurement quality. On the other hand, an RMSEA value of 0.20, traditionally suggesting poor fit, can still be an indicator of acceptable fit and good measurement quality, especially when sample sizes are fairly large ($N > 1000$). Recent studies (e.g., McNeish & Wolf, 2021; Niemand & Mai, 2018) indicate that rigid cutoffs can become imprecise, and, consequently, using flexible or dynamic cutoff values would be more appropriate.

Therefore, it is inadequate to conclude that a scale is invalid because of weak model fit determined by the widely used fit indices. Construct and predictive validity should be considered. Scales can perform better by reducing their size and/or complexity. This approach,

however, may come at cost of construct and predictive validity (Hopwood & Donnellan, 2010). Marsh et al. (2004) explained that overgeneralizing the golden rules is not an adequate approach. Authors commented that these cutoff values are "based largely on intuition and have little theoretical justification" (p. 321). Their limitations should be acknowledged, not blindly accepted.

The next analytical step of our plan was to identify items that contribute to poor model fit or behave differently across samples; these items are known to have differential item functioning (DIF; Holland & Thayer, 1986). Various procedures have been proposed to identify DIF items. Namely, researchers have suggested utilizing an ordinal or logistic regression approach based on item response theory (Fischer & Karl, 2019; Swaminathan & Rogers, 1990). However, one limitation to this approach is that it does not control for potential confounding variables (such as indices of acquiescence or extreme responding). Therefore, we selected a structural equation modeling approach that allows us to include such indices when identifying DIF items: the multiple indicator, multiple cause (MIMIC) model (Jöreskog & Goldberger, 1975). In this approach, researchers include dummy-coded predictor variables that indicate group membership. They then compare models in which the grouping variables are set to predict each latent variable (factor) within the model vs. those in which the grouping variables additionally and successively predict each item (i.e., constrained baseline approach; Wang, 2004). Potential DIF items are identified based on changes in model fit as well as regression coefficients. For example, if multiple grouping variables significantly differ from the reference group on a particular item, then that item is one with DIF.

Once DIF items were identified for the scale, we ran modified multigroup CFA models to test if model fit improves (compared to the original multigroup CFA model) when removing the non-invariant items. Doing so allowed us to identify a cross-culturally invariant version of QLS.

Finally, as an additional test of the model fit of our theoretical model, we sought to compare this model to one in which all items are permitted to load on each latent variable. To do so, we utilized a multigroup exploratory structural equation modeling approach, which combines procedures from exploratory factor analysis (EFA) and structural equation modeling (SEM), namely multigroup confirmatory factor analysis (Fischer & Karl, 2019; Perry et al., 2015). In essence, this procedure allowed us to estimate the factor loadings of each item in our scale, using EFA, on a preset number of factors (in our case, we will choose 4, which reflects the number of theoretical factors in our model). We then inputted these specified factor loadings for each of the four factors in a multigroup

confirmatory factor analysis model to test configural invariance. We then tested two models: one in which indices of acquiescence and extreme responding were not included as controls and one in which they were included. Theoretically, this ESEM model approach should result in better model fit, as the factor loading estimates should be most accurate for all items across all latent variables (Fischer & Karl, 2019; Perry et al., 2015). However, if the model fit for these models is similar to or no worse than our theoretical configural models (in which each item is only permitted to load on one latent variable), we have additional evidence supporting the goodness of fit of our theoretical model.

Quality criteria
Results
Multigroup CFA I: initial configural models

Baseline model (not controlling for acquiescence and extreme responding)   We first tested the invariance of our scale across our eight cultural samples using a multigroup CFA approach. We were particularly interested in measuring invariance of a baseline configural model, which allows all factor loadings, intercepts, and variances to vary across cultural samples. Poor model fit would indicate that (1) the scale has a different structure of measurement for different groups, or (2) certain items perform differently across the cultural samples. Therefore, different from the traditional approach (Cheung & Rensvold, 1999; Raju et al., 2002; Vandenberg & Lance, 2000), we did not compare this baseline model to more restricted models.

In our case, we used a maximum likelihood estimator with robust standard errors and a corrected χ^2 test of global fit based on the Yuan-Bentler test statistic (Yuan & Bentler, 1998). This is the most appropriate estimator as it accounts for missing and nonnormal data (Muthén & Muthén, 2017; Yuan & Bentler, 1998). Moreover, we wanted to estimate all factor loadings, so we constrained the latent variable variances to one. We followed Hu and Bentler's (1998, 1999) suggestion to testing model fit. We first measured absolute or global fit using the Yuan-Bentler corrected χ^2 test (Yuan & Bentler, 1998). However, the size of the correlations in the model affects χ^2, such that larger correlations indicate poor fit. Because of this, χ^2 tests of global fit were oftentimes significant ($p < .05$), indicating poor model fit. Therefore, we also tested model fit using three additional measures. We assessed another measure of global or absolute fit, standardized root-mean-square residual (SRMR), which is not as prone to becoming inflated due to degrees of freedom and sample size. We also tested incremental fit using

Table 2 Multigroup CFA/SEM measurement invariance configural model fit indices

Model	Description	χ^2	BIC	df	CFI	RMSEA [90% *CI*]	SRMR
1	Baseline	17133.907	281377.878	6606	0.817	.088 [.086, .090]	.064
2	Controlling for response sets[a]	18778.581	272117.871	7254	0.828	.087 [.086, .089]	.062
3	Removing Turkey sample	16856.649	256963.258	6448	0.834	.088 [.086, .089]	.060
4	Removing Turkey and Iran samples	14605.926	228251.704	5642	0.840	.088 [.086, .089]	.057
5	Removing Turkey, Iran, and Midwest samples	12712.719	214411.701	4836	0.844	.086 [.084, .088]	.057
6	Removing Turkey, Iran, Midwest, and Brazil samples	10653.304	182703.263	4030	0.845	.086 [.084, .087]	.056
7	Removing items with ≥ 6 sig. differences[b]	13040.926	225428.550	4923	0.849	.090 [.088, .091]	.062
8	Removing items with ≥ 5 sig. differences[b]	10374.024	203776.410	4059	0.866	.088 [.086, .090]	.059
9	Model 2 including 2 higher-order latent factors[c]	18662.204	272474.078	7209	0.828	.087 [.086, .089]	.062
10	Model 7 including 2 higher-order latent factors[c]	12921.780	225784.937	4878	0.849	.090 [.088, .092]	.062
11	Model 8 including 2 higher-order latent factors[c]	10259.014	204132.863	4014	0.866	.089 [.087, .091]	.059
12	ESEM model not controlling for response sets	16983.637	278909.593	6966	0.822	.085 [.083, .086]	0.451
13	ESEM model controlling for response sets	19546.090	270829.735	7614	0.816	.088 [.086, .089]	0.346

[a] Models 3–11 and model 13 control for response sets (acquiescence and extreme responding). [b] These significant differences were determined based on MIMIC model significant regression coefficients across dummy-coded grouping variables (see Table 1 of Supplementary Materials). [c] The two higher-order latent factors are (1) feelings towards the partner and (2) feelings about the relationship. χ^2 = robust χ^2 test of global fit (equivalent to Yuan-Bentler corrected test statistic). Measures of CFI and RMSEA are based on robust estimates for nonnormal data (Brosseau-Liard et al., 2012; Brosseau-Liard & Savalei, 2014)

the comparative fit index (CFI), as well as parsimonious fit using the root-mean-square error of approximation (RMSEA; Hu & Bentler, 1998, 1999). We compared each of the three indices of model fit to the good model fit "rules of thumb" (Hu & Bentler, 1999): SRMR ≤ .08, CFI ≥ 0.95 (or 0.90), and RMSEA ≤ .06 or .08. Our measures of CFI and RMSEA were based on robust estimates outlined by Brosseau-Liard and Savalei (2014) and Brosseau-Liard et al. (2012), respectively. Like the robust maximum likelihood estimator, these estimates also account for the nonnormality of our data[1].

As the results in Table 2 show (see model 1), this baseline configural model exhibited relatively poor model fit in terms of indices of CFI and RMSEA), even though the indices are not too far from a conventional cutoff. This suggests, as expected, that our scale is noninvariant across our eight cultural samples. Keep in mind that this model does not control for indices of acquiescence and extreme responding, two response sets that are common in multicultural samples. Therefore, we turned to addressing this issue next.

Response set model (controlling for acquiescence and extreme responding) As mentioned in the "Method" section, we created indices of acquiescence and extreme responding. Therefore, we tested the multigroup model fit of the scale while controlling for these indices.

Specifically, within each model, we regressed each latent variable on both of our response set indices. As the model fit indices presented in Table 2 indicate (see model 2), this new model, in which acquiescence and extreme responding were controlled, provided a better fitting model compared to our initial baseline model. Therefore, for all remaining models presented below, we included acquiescence and extreme responding as predictors of our latent (factor) variables. Nevertheless, in terms of meeting the criteria outlined by Hu and Bentler (1999), it is a little below the cutoff for good model fit. This suggests we need to investigate how the cultural samples under study contribute to this poor fitting model.

Individual group CFAs
We also conducted CFAs on each cultural sample individually. As the results of these analyses show (see Table 3), the model fit was adequate for most of the individual cultural samples. However, the data from four samples — Turkey, Iran, Brazil, and Midwest of the USA — exhibited relatively lower model fit than other samples. Although this may suggest that our scale may perform differently within these samples compared to the others, we must also consider sample sizes of each sample. Therefore, we turned to analyzing adjusted models while taking these sample differences into consideration.

Multigroup CFA II: configural model, removing some samples
As stated above, the Turkey, Iran, Midwest of the USA, and Brazil samples had a relatively lower model fit than the rest of the cultural samples. Therefore, to test how

---

[1] We use the same model fit criteria and robust estimates of global fit (Yuan-Bentler corrected χ^2), CFI, and RMSEA for all successive CFA/SEM models.

Table 3 Individual group CFA model fit indices for each cultural sample

Model	Cultural sample	χ^2	BIC	df	CFI	RMSEA [90% *CI*]	SRMR
1	Brazil	2060.865	31331.848	806	0.833	.089 [.085, .094]	.065
2	Iran	2277.181	28292.263	806	0.770	.089 [.085, .094]	.080
3	Portugal	2076.252	27139.002	806	0.826	.091 [.086, .096]	.067
4	Saint Petersburg, Russia	2099.277	37969.917	806	0.858	.084 [.080, .089]	.049
5	Tyumen, Russia	2184.895	39652.294	806	0.835	.086 [.082, .091]	.056
6	Hawaii, USA	2024.954	27735.398	806	0.834	.089 [.084, .094]	.060
7	Midwest, USA	1840.879	13401.962	806	0.793	0.106 [0.100, 0.113]	.057
8	Southeast, USA	2251.132	49074.915	806	0.862	.080 [.076, .084]	.052
9	Turkey	4881.180	14717.393	806	0.729	.082 [.077, .087]	.084

χ^2 = robust χ^2 test of global fit (equivalent to Yuan-Bentler corrected test statistic). Measures of CFI and RMSEA are based on robust estimates for nonnormal data (Brosseau-Liard et al., 2012; Brosseau-Liard & Savalei, 2014). All models control for acquiescence and extreme responding

much these samples decrease an overall model fit of our multiple group invariance test, we ran three additional multigroup CFAs, with each cultural sample successively omitted from analysis. First, we removed the Turkey sample. As the results in Table 2 (model 3) show, this model exhibited improved model fit. We then additionally removed the Iran sample from analysis and tested the model fit across the remaining samples (model 4). Removing the Iran sample also resulted in improved model fit compared to the previous model. Third, we removed the Midwest of the USA sample and consequently tested the model fit across the remaining samples (model 5). Finally, we removed the Brazil sample from analysis and again tested the revised model's model fit (model 6). Removing the Brazil sample resulted in an even more improved model compared to the previous two models and the configural model. The model fit became reasonably acceptable — a little below the conventional cutoff.

Thus, we can conclude that more samples in cross-cultural analysis may cause an increased likelihood of variance and poorer model fit. Therefore, by eliminating some samples from analysis, researchers can increase invariance and consequently validity of cross-cultural comparison.

Another possible approach to increase invariance of a scale in cross-cultural comparison is elimination of variant items. By doing so, invariance of measurement across all cultural samples can be reached with development of a cross-culturally acceptable model and a shorter version of the scale. To test this approach, we opted to include all samples in all successive models. So, we turned to identifying cross-culturally variant items (i.e., items with DIF).

DIF detection: MIMIC models

To identify cross-culturally variant items—the items with DIF—we employed the MIMIC model approach.

One major advantage of this approach is that it allows us to control for our indices of acquiescence and extreme responding. This approach is traditionally used to compare the performance of two groups on scale items (Finch, 2005; Kim et al., 2012; Woods, 2009) but can also be done for more than two groups (Chun et al., 2016). To create variables indicating group membership for our eight cultural samples, we created seven dummy-coded variables. To do so, we needed to determine a reference group. Based on it having arguably the best model fit (especially in terms of RMSEA; see Table 3) as well as having the largest sample of all cultural samples, we selected the Southeast of the USA sample as our reference group.

For our MIMIC models, we used the constrained baseline approach (Wang, 2004). In this approach, we created a constrained baseline, in which each latent factor is regressed on each grouping variable. We then compared this model to models in which each item is also regressed on each grouping variable. This results in a total of 41 models (one for the grouping variable and one for each of the 40 items in our scale). This approach allowed us to compare performance between each group and the reference group for each individual item within our scale. Therefore, in addition to assessing results of model fit, we were also interested in the regression coefficients for each dummy-coded grouping variable. If a regression coefficient between a cultural sample and the reference group (the Southeast sample) is significant, this indicates a significant difference between the two groups and, consequently, potential DIF for that item. Moreover, more significant regression coefficients across cultural samples (when compared to the Southeast of the USA sample) for a single item indicate a greater level of DIF for that item. We present the model fit statistics as well as the regression coefficients for each grouping variable for all 41 models in Table 2 of Supplementary materials. Using

these findings, particularly the regression coefficients, we were able to determine several items with various degree of DIF, which we decided to account for in our adjusted multigroup CFA invariance models. We discuss these items and the adjusted models below.

Multigroup CFA III: configural models, removing DIF items

Based on the results from the MIMIC models, particularly the regression coefficients, we identified several items with varying degrees of differential functioning (the candidates for high DIF across samples). We focused on the items (a) with six or more significant regression coefficients across the seven dummy-coded grouping variables and (b) those with five or more significant regression coefficients.

We identified seven items with six or more significant regression coefficients. Removing these items left us with a 33-item revised scale (9 items for compassion, 7 items for affection, 9 items for closeness, and 8 items for commitment). We then ran another multigroup CFA to test measurement invariance on this revised measure (see model 7 of Table 2). This model showed improved model fit with indices slightly below conventional cutoff. See Table 3 of Supplementary materials for the removed items compared to the original 40-item measure.

We then identified 3 additional items that have four or more significant regression coefficients and removed them in addition to the previously removed items to test this new version of the scale. This left us with a 30-item measure (7 items for compassion, 7 items for affection, 8 items for closeness, and 8 items for commitment). We then ran another multigroup CFA to test measurement invariance on this shorter version of scale (see model 8 of Table 2). This 30-item model provided us with the best fitting model, especially in terms of CFI. This analysis provides further support that (1) DIF can serve as an indicator of cross-culturally similar or culturally different meanings of scale items, and (2) removing those items with high DIF allows for creating shorter and more cross-culturally universal versions of a scale with adequate fit model with acceptable conventional cutoff. See Table 3 of Supplementary materials for the removed items compared to the original 40-item measure.

Multigroup CFA IV: configural models, accounting for two higher-order factors

The QLS also includes two higher-order factors: (1) feelings toward the partner and (2) feelings about the relationship. The feelings of compassion and affection theoretically belong to the feelings toward the partner higher-order factor, while the feelings of closeness and commitment theoretically belong to the feelings about the relationship higher-order factor (Karandashev

& Evans, 2019). As such, we tested the same configural model (controlling for acquiescence and extreme responding) with these higher-order latent variables included in the model. As the results from Table 2 (model 9) suggest, this model shows the same acceptable model fit as the four-factor model (controlling for acquiescence and extreme responding) (Table 2, model 2). Additionally, we included the two higher-order factors in the models in which we removed our identified DIF items: those with 5 or more significant regression coefficients from the MIMIC models (Table 2, model 10) and those with 4 or more significant regression coefficients (Table 2, model 11). These models also showed the same acceptable model fit as their non-higher-order factor counterparts (models 7 and 8 of Table 2, respectively). This suggests that the theoretical model (Karandashev & Evans, 2019) is supported, particularly when accounting for items that differ across cultural samples.

Multigroup CFA V: exploratory structural equation modeling (ESEM)

Finally, we compared our multigroup CFA models to a model in which all items are permitted to load on the four theoretical factors. Doing so allowed us to understand how well our theoretical model—in which each factor is only permitted to load on one latent variable or factor—holds when pitted against a model in which all items are permitted to load on each factor based on predetermined factor loadings. Therefore, we employed a multigroup exploratory structural equation modeling approach. This approach combines procedures from exploratory factor analysis (EFA) and structural equation modeling (SEM), namely multigroup confirmatory factor analysis (Fischer & Karl, 2019; Perry et al., 2015).

In line with the analytic procedure outlined by Fischer and Karl (2019), we first ran an EFA on the entire sample to estimate the factor loadings of all items on four theoretical factors (which mirrors the number of factors in our theoretical model). For this EFA, we used oblique oblimin rotation and a maximum likelihood estimation for the factor method. The resulting factor loadings of all items on each of the latent variables were specified in a multigroup confirmatory factor analysis to test configural invariance. We ran these configural invariance tests for a model in which indices of acquiescence and extreme responding were not included as controls (see Table 2, model 12) and one in which these indices were included as controls (see Table 2, model 13). As the results suggest, both ESEM models exhibit model fit similar to our theoretical model, yet with lower indices compared to other methods of analysis. Therefore, models in which each

item is only permitted to load on one of our theoretical latent variables are sufficient, and they fall in line with our theoretical structure of love attitudes.

Discussion

The aim of the study was to (1) explore the possible methods for statistical analysis of psychometric scales with the data obtained in several cultural samples and (2) apply these methods to the investigation of a comprehensive set of constructs associated with love. The main innovation of this methodology was a shift from a traditional imposed-etic approach to the derived-etic approach, at least at the stage of data analysis.

In the traditional imposed-etic approach, researchers modify and adjust items of scale for culturally different samples to emulate a factor structure of an original reference culture. They do this on the stage of scale adaptation and validation. This approach has been adequate in case of theoretically imposed models of love (e.g., Triangular Love Scale, Sorokowski et al., 2021; Sternberg, 1997), where the higher-order constructs/factors have been first theoretically postulated and then second their descriptive items compiled. Therefore, the specific text and meaning of the items is of secondary importance if they load on the same higher-order factor. Even removing some items keeps a construct/factor intact and adequate. This approach gives a lot of freedom in adaptation of original scale in other culture — the items can deviate from original text substantially.

In the derived-etic approach, which we have proposed, researchers do not strive to reproduce and validate the same scale as in the culture of origin. Instead, they attempt to identify which part of the scale is cross-culturally similar (and can be used for cross-cultural comparison) and which part is cross-culturally variable (and cannot be compared in the context of a current study). The latter items can be valuable for other analyses. Thus, researchers adapt and validate a scale on the stage of data analysis.

This new approach is important in cases when researchers first compile descriptive basic constructs/items (each with distinct and valuable meaning) and then try to identify their higher-order structure in terms of first- and second-order constructs/factors. Therefore, the specific text and meaning of each item is of primary importance.

Researchers can also employ this new approach with theoretically imposed models of love that allow eliminating items with high DIF items from a model and creating a revised short version, which will be more adequate for cross-cultural comparison.

Conclusions, limitations, and recommendations

The results of our study first allow us to control for scores for acquiescence and extreme biases in cross-cultural psychometric research. This improves model fit of a scale and makes the scale more comparable across samples. As our results show, along with other studies (Bachman & O'Malley, 1984; van Herk et al., 2004), this improvement is not substantial, yet it adds accuracy to measurement. Based on the results of our study, we recommend using the approach described by Bachman and O'Malley (1984) as well as van Herk et al. (2004) to compute indices of acquiescence and extreme responding. Including these as predictor variables in tests of model fit using CFA/SEM allows researchers to control for any variance for which these variables account.

When there are fewer groups or samples included in measurement invariance tests, the likelihood of cross-cultural invariance increases. The increasing number of cultural samples in MGCFA increases the likelihood of non-invariance across samples. However, some cultural samples add more variance than others. Researchers should be cautious when comparing model fit results to the traditional "rules of thumb" specified by Hu and Bentler (1999). Indeed, researchers have challenged this notion of having such strict cutoff criteria to determine if the model being tested indeed fits the data (Perry et al., 2015). Therefore, determining whether a tested model has "good" or "poor" model fit should not be solely based on whether or not each model fit index meets its corresponding "rule of thumb."

Several studies demonstrated difficulties in achieving an acceptable CFA model fit on personality scales (e.g., Hopwood & Donnellan, 2010; Marsh et al., 2010) and sport and exercise psychology scales (Perry et al., 2015). Hopwood and Donnellan (2010) explored eight widely utilized personality scale measurements and found that none of these scales came close to these recommended cutoff values. Even the best-performing scale, despite being recognized as an appropriate evaluation of personality, reached a model fit lower than the usually accepted criteria. Authors (Hopwood & Donnellan, 2010) contend that this model misfit suggests that using CFA cutoff values is inadequate for multidimensional measures. They recommend that researchers should be cautious when they interpret the results of the CFA. These cutoff values, which are frequently applied for the fit indices, are unrealistic to achieve for most measures. Therefore, assessing factorial validity and finding acceptable levels of fit are not straightforward. These rigid cutoff values should not be strict standards for interpretation (Hopwood & Donnellan, 2010).

It is very challenging to expect cross-cultural universality of a scale. Based on the results of our study, we

recommend testing the model fit of both measurement invariance models, including all samples, as well as testing the model fit for each individual group. Doing so would allow researchers to understand for which groups a scale performs well or poorly. This knowledge may then allow researchers to assess whether there may be sample-specific characteristics or scale-specific characteristics (e.g., translation issues) that contribute to poor model fit.

In case of cross-cultural variance, identification of items with high DIF allows for creating a revised shorter version of a scale by dropping these items. These results in a scale more adequate for cross-cultural comparison. Based on the results of our study, we recommend testing DIF for each scale item and removing items that have high DIF. Based on the analytic procedure used to detect DIF (e.g., a logistic regression approach or a MIMIC model approach), the criteria for DIF items should be up to the researcher's discretion.

All these methods, which are summarized above, are successfully applied in the revisions of QLS. Based on these analyses, the shortened versions of the scale (see Table 2 of Supplementary materials) are recommended for cross-cultural comparisons if researchers are interested in comparison of the first-order factors (compassion, affection, closeness, commitment) and second-order factors (feelings toward the partner and feelings about the relationship).

If researchers are interested in comparison of basic love constructs, which are expressed in single items, then comparison of only those items with low DIF is adequate. Several studies have shown that the single-item dimensions are as valid as multi-item dimensions (e.g., Bartholomew, 1994; Robins et al., 2001). Researchers view single-item scales as psychometrically valid options, comparable to longer multi-item scales, which can be redundant (Barrett & Paltiel, 1996).

Limitations

Although our analytic procedure to test cross-cultural invariance of the QLS yielded an adequately fitting model, several limitations must be taken into consideration for future work involving cross-cultural validation of love scales.

First, the number of cultural samples and methods which are used in the analyses reported in this article is limited. Assessing the QLS within an increased number of cultural samples would shed more light on how this measure can be implemented in different cultural contexts.

Second, the results have demonstrated that increasing the number of cultural samples increases the likelihood of measurement invariance of a love scale. We don't believe in cross-culturally universal scales, which, once

validated, would be valid for any other studies and samples. Invariance is the function not only of the scale itself but also of samples under study. Such exploration is necessary in any cross-cultural study to verify that it is valid to compare samples.

Third, the analyses reported in this article apply only several approaches of many potential statistical analyses to test cross-cultural invariance. Other methods can be also productive in the future studies. For example, ant colony optimization (ACO) can be used to select items based on both uniform and nonuniform DIF and, consequently, create shortened versions of scales with better model fit (Olaru & Danner, 2020). Other methods, which can be used in the future studies, include alignment optimization (Asparouhov & Muthén, 2014) and Bayesian approximate invariance (Cieciuch et al., 2014; Muthén & Asparouhov, 2012).

Recommendations

We recommend that the data from future samples measuring the QLS using both previously developed language versions and new language versions should be analyzed guided by the following steps.

1. Compute indices of acquiescence and extreme responding to include as controls when carrying out all subsequent analyses. This will allow the researcher to measure and control the variance accounted for by individual response styles.
2. Using confirmatory factor analysis (CFA), test the model fit of the full 40-item model as well as the revised 33- and 30-item models. These model fit comparisons would provide evidence for or against the shortened versions of the QLS.
3. In case of comparisons across cultural samples, check measurement invariance of the scale across samples using MG-CFA and measures of DIF as described in this article for 40-, 31-, and 24-item models. Choose which one has adequate model fit.

Supplementary Information

> **Additional file 1: Table 1**. Means, Standard Deviations, and Intercorrelations of Composite QLS Factors and QLS Items. **Table 2**. Constrained Baseline MIMIC Models for Each Item. **Table 3**. Original Quadrangular Love Scale and the Items Retained or Removed for the Two Shortened Versions.

Authors' contributions
Victor Karandashev – literature review, research design, data collection, analyses, and writing. Nicholas D. Evans – literature review, research design, data collection, analyses, and writing, Félix Neto – Portuguese version and

data collection, Elena Zarubko – Russian version and data collection, Veronika Artemeva – data collection, Sadeq Fallah – Iranian version and data collection, Vicente Cassepp-Borges and Eliany Nazaré Oliveira –Brazilian version and data collection in Brazil, Havvanur Akyol - Turkish version and data collection, Duygu Dincer – Turkish version and data collection. The author(s) read and approved the final manuscript.

Declarations

Ethics approval and consent to participate
Yes, the study has been approved by the Institutional Review Board of Aquinas College, and participants gave an informed consent to participate.

Consent for publication
The article contains no individual person's data in any identifiable form.

Competing interests
The authors declares that they have no competing interests.

Author details
[1]Aquinas College, Grand Rapids, USA. [2]University of Texas at El Paso, El Paso, USA. [3]Universidade do Porto, Porto, Portugal. [4]Tyumen State University, Tyumen, Russia. [5]Baltic Academy of Education, Saint Petersburg, Russia. [6]Tarbiat Modares University, Tehran, Iran. [7]Universidade Federal Fluminense, Volta Redonda, RJ, Brazil. [8]Universidade Estadual Vale do Acaraú, Sobral, Ceará, Brazil. [9]Marmara University, Istanbul, Tukey. [10]Biruni University, Istanbul, Turkey.

References
Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(4), 495–508.

Bachman, J. G., & O'Malley, P. M. (1984). Yea-saying, nay-saying, and going to extremes: Black-white differences in response styles. *Public Opinion Quarterly*, *48*, 491–509.

Barrett, P. T., & Paltiel, L. (1996). Can a single item replace an entire scale? POP vs. the OPQ 5.2. *Selection and Development Review*, *12*, 1–4.

Bartholomew, K. (1994). Assessment of individual differences in adult attachment. *Psychological Inquiry*, *5*, 23–27.

Boer, D., Hanke, K., & He, J. (2018). On detecting systematic measurement error in cross-cultural research: A review and critical reflection on equivalence and invariance tests. *Journal of Cross-Cultural Psychology*, *49*(5), 713–734.

Brosseau-Liard, P. E., & Savalei, V. (2014). Adjusting incremental fit indices for nonnormality. *Multivariate Behavioral Research*, *49*(5), 460–470. https://doi.org/10.1080/00273171.2014.933697.

Brosseau-Liard, P. E., Savalei, V., & Li, L. (2012). An investigation of the sample performance of two nonnormality corrections for RMSEA. *Multivariate Behavioral Research*, *47*(6), 904–930. https://doi.org/10.1080/00273171.2012.715252.

Byrne, B. M., & Van de Vijver, F. J. R. (2017). The maximum likelihood alignment approach to testing for approximate measurement invariance: A paradigmatic cross-cultural application. *Psicothema*, *29*, 539–551. https://doi.org/10.7334/psicothema2017.178.

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology*, *95*(5), 1005–1018. https://doi.org/10.1037/a0013193.

Cheung, G. W., & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization an proposed new method. *Journal of Management*, *25*(1), 1–27. https://doi.org/10.1016/S0149-2063(99)80001-4.

Chun, S., Stark, S., Kim, E. S., & Chernyshenko, O. S. (2016). MIMIC methods for detecting DIF among multiple groups: Exploring a new sequential-free baseline procedure. *Applied Psychological Measurement*, *40*(7), 486–499. https://doi.org/10.1177/0146621616659738.

Cieciuch, J., Davidov, E., Schmidt, P., Algesheimer, R., & Schwartz, S. H. (2014). Comparing results of an exact vs. an approximate (Bayesian) measurement invariance test: A cross-country illustration with a scale to measure 19 human values. *Quantitative Psychology and Measurement*, *5*, 982. https://doi.org/10.3389/fpsyg.2014.00982.

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross-national research. *Annual Review of Sociology*, *40*, 55–75.

Fehr, B. (1994). Prototype-based assessment of laypeople's views of love. *Personal Relationships*, *1*, 309–331.

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement*, *29*(4), 278–295. https://doi.org/10.1177/0146621605275728.

Fischer, R., & Karl, J. A. (2019). A primer to (cross-cultural) multi-group invariance testing possibilities in R. *Frontiers in psychology*, *10*, 1507 https://www.frontiersin.org/articles/10.3389/fpsyg.2019.01507/full.

Graham, J. M. (2011). Measuring love in romantic relationships: A meta-analysis. *Journal of Social and Personal Relationships*, *28*, 748–771. https://doi.org/10.1177/0265407510389126.

Hambleton, R., & Zenisky, A. (2010). Translating and adapting tests for cross-cultural assessments. In D. Matsumoto, & F. Vijver (Eds.), *Cross-cultural research methods in psychology*, (pp. 46–74). Cambridge University Press.

He, J., & van de Vijver, F. (2012). Bias and equivalence in cross-cultural research. *Online Readings in Psychology and Culture*, *2*(2). https://doi.org/10.9707/2307-0919.1111.

Hirschfeld, G., & Von Brachel, R. (2014). Improving multiple-group confirmatory factor analysis in R–A tutorial in measurement invariance with continuous and ordinal indicators. *Practical Assessment, Research, and Evaluation*, *19*(1), 7.

Holland, P., & Thayer, D. T. (1986). Differential item functioning and the Mantel-Haenszel procedure. *ETS Research Report Series*, *1986*(2), i–24. https://doi.org/10.1002/j.2330-8516.1986.tb00186.x.

Hopwood, C. J., & Donnellan, M. B. (2010). How should the internal structure of personality inventories be evaluated? *Personality and Social Psychology Review*, *14*, 332–346. https://doi.org/10.1177/1088868310361240.

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*(4), 424–453. https://doi.org/10.1037/1082-989X.3.4.424.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118.

Hui, C. H., & Triandis, H. C. (1985). Measurement in cross-cultural psychology: A review and comparison of strategies. *Journal of Cross-cultural Psychology*, *16*(2), 131–152.

Johnson, T. P. (1998). Approaches to equivalence in cross-cultural and cross-national survey research. In J. A. Harkness (Ed.), *Cross-cultural survey equivalence*, (pp. 1–40). Zentrum für Umfragen, Methoden und Analysen.

Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of American Statistical Association*, *70*(351a), 631–639. https://doi.org/10.1080/01621459.1975.10482485.

Karandashev, V. (2017). *Romantic love in cultural contexts*. Springer International Publishing.

Karandashev, V. (2019). *Cross-cultural perspectives on the experience and expression of love*. Springer International Publishing.

Karandashev, V. (2021). Cultural diversity of romantic love experience. In C. Mayer & E. Vanderheiden (Eds.), International handbook of love: Transcultural and transdisciplinary perspectives (pp. 59-79). Springer. https://doi.org/10.1007/978-3-030-45996-3.

Karandashev, V., & Evans, N. D. (2019). Quadrangular love theory and scale: Validation and psychometric investigation. *Journal of Methods and Measurement in the Social Sciences*, *10*(1), 1–35.

Kenny, D. A. (2012). Measuring model fit. Retrieved from http://davidakenny.net/cm/fit.htm

Kim, E. S., Yoon, M., & Lee, T. (2012). Testing measurement invariance using MIMIC: Likelihood ratio test with a critical value adjustment. *Educational and Psychological Measurement*, *72*(3), 469–492. https://doi.org/10.1177/0013164411427395.

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis testing approaches to setting cutoff values for fit indexes and dangers in overgeneralising Hu & Bentler's (1999) findings. *Structural Equation Modeling*, *11*, 320–341. https://doi.org/10.1207/s15328007sem1103_2.

Marsh, H. W., Ludtke, O., Muthen, B., Asparouhov, T., Morin, A. J. S., Trautwein, U., & Nagengast, B. (2010). A new look at the big-five factor structure through exploratory structural equation modeling. *Psychological Assessment*, *22*, 471–491. https://doi.org/10.1037/a0019227.

Masuda, M. (2003). Meta-analyses of love scales: Do various love scales measure the same psychological constructs? *Japanese Psychological Research*, *45*, 25–37.

McNeish, D., An, J., & Hancock, G. R. (2018). The thorny relation between measurement quality and fit index cutoffs in latent variable models. *Journal of Personality Assessment*, *100*(1), 43–52.

McNeish, D., & Wolf, M. G. (2021). Dynamic fit index cutoffs for confirmatory factor analysis models. In *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000425.

Milfont, T. L., & Fischer, R. (2010). Testing measurement invariance across groups: Applications in cross-cultural research. *International Journal of Psychological Research*, *3*(1), 111–130.

Morren, M., Gelissen, J., & Vermunt, J. (2012). The impact of controlling for extreme responding on measurement equivalence in cross-cultural research. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *8*(4), 159–170. https://doi.org/10.1027/1614-2241/a000048.

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods*, *17*, 313–335. https://doi.org/10.1037/a0026802.

Muthén, L. K., & Muthén, B. O. (2017). *Mplus User's Guide*, (8th ed.,). Muthén & Muthén.

Niemand, T., & Mai, R. (2018). Flexible cutoff values for fit indices in the evaluation of structural equation models. *Journal of the Academy of Marketing Science*, *46*(6), 1148–1172.

Olaru, G., & Danner, D. (2020). Developing cross-cultural short scales using ant colony optimization. *Assessment*, *28*(1), 199–200. https://doi.org/10.1177/1073191120918026.

Perry, J. L., Nicholls, A. R., Clough, P. J., & Crust, L. (2015). Assessing model fit: Caveats and recommendations for confirmatory factor analysis and exploratory structural equation modeling. *Measurement in Physical Education and Exercise Science*, *19*, 12–21. https://doi.org/10.1080/1091367X.2014.952370.

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology*, *87*(3), 517–529. https://doi.org/10.1037/0021-9010.87.3.517.

Robins, R. W., Hendin, H. M., & Trzesniewski, K. H. (2001). Measuring global self-esteem: Construct validation of a single-item measure and the Rosenberg self-esteem scale. *Personality and Social Psychology Bulletin*, *27*(2), 151–161.

Sorokowski, P., Sorokowska, A., Karwowski, M., Groyecka, A., Aavik, T., Akello, G., … Sternberg, R. J. (2021). Universality of the triangular theory of love: Adaptation and psychometric properties of the triangular love scale in 25 countries. *The Journal of Sex Research*, *58*(1), 106–115.

Steenkamp, J. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*, 78–90.

Stein, J. A., Lee, J. W., & Jones, P. S. (2006). Assessing cross-cultural differences through the use of multiple-group invariance analyses. *Journal of Personality Assessment*, *87*(3), 249–258. https://doi.org/10.1207/s15327752jpa8703_05.

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

Swaminathan, H., & Rogers, J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361–370. https://doi.org/10.1111/j.1745-3984.1990.tb00754.x.

Van De Schoot, R., Schmidt, P., De Beuckelaer, A., Lek, K., & Zondervan-Zwijnenburg, M. (2015). Measurement invariance. *Frontiers in Psychology*, *6*, 1064.

Van de Vijver, F. J. R., & Leung, K. (1997). *Methods and data-analysis for cross-cultural research*. Sage.

Van de Vijver, F. J. R., & Leung, K. (2011). Equivalence and bias: A review of concepts, models, and data analytic procedures. In D. Matsumoto, & F. J. R. Van de Vijver (Eds.), *Cross-cultural research methods in psychology*, (pp. 17–45). Cambridge University Press.

van Herk, H., Poortinga, Y. H., & Verhallen, T. M. M. (2004). Response styles in rating scales: Evidence of method bias in data from 6 EU countries. *Journal of Cross-Cultural Psychology*, *35*(3), 346–360. https://doi.org/10.1177/0022022104264126.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurements invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, 4–69.

Wang, W.-C. (2004). Effects of anchor item methods on the detection of differential item functioning within the family of Rasch models. *Journal of Experimental Education*, *72*(3), 221–261. https://doi.org/10.3200/JEXE.72.3.221-261.

Woods, C. M. (2009). Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. *Multivariate Behavioral Research*, *44*(1), 1–27. https://doi.org/10.1080/00273170802620121.

Yuan, K., & Bentler, P. M. (1998). Robust mean and covariance structure analysis. *British Journal of Mathematical and Statistical Psychology*, *51*(1), 63–88. https://doi.org/10.1111/j.2044-8317.1998.tb00667.x.

Publisher's Note