Published in final edited form as:
Nurs Res. 2012 May ; 61(3): 171–180. doi:10.1097/NNR.0b013e3182544750.

Establishing Measurement Invariance: English and Spanish Paediatric Asthma Quality of Life Questionnaire

Karen H. Sousa, PhD, RN, FAAN[Associate Dean for Research and Extramural Affairs and Professor],
University of Colorado College of Nursing Aurora, Colorado

Stephen G. West, PhD[Professor Psychology],
Arizona State University Phoenix, Arizona

Stephanie E. Moser, PhD,
Veteran's Administration Medical Center Ann Arbor, Michigan

Judy A. Harris, MS, RN, CPNP[Director], and
Breathmobile Phoenix Children's Hospital Phoenix, Arizona

Susanne W. Cook, PhD, RN[Retired]

Abstract

Background—Registered nurses and nurse researchers often use questionnaires to measure patient outcomes. When questionnaires or other multiple item instruments have been developed using a relatively homogeneous sample, the suitability of even a psychometrically well-developed instrument for the new population comes into question. Bias or lack of equivalence can be introduced into instruments through differences in perceptions of the meaning of the measured items, constructs, or both, in the two groups.

Objective—To explain measurement invariance and illustrate how it can be tested using the English and Spanish versions of the Paediatric Asthma Quality of Life Questionnaire (PAQLQ).

Method—A sample of 607 children from the Phoenix Children's Hospital Breathmobile was selected for this analysis. Ages were 6 to 18 years in age; 61.2% completed the PAQLQ in Spanish. Testing measurement invariance using multiple group confirmatory factor analysis, a series of hierarchical nested models, is demonstrated. In assessing the adequacy of the fit of each model at each stage, both $\chi 2$ tests and goodness-of-fit indexes were used.

Results—The test of measurement invariance for the one-factor model showed that the English and Spanish versions of the scale met the criteria for measurement invariance. The level of strict invariance (equal factor loadings, intercepts, and residual variances between groups) was achieved.

Discussion—Confirmatory factor analysis is used to evaluate the structural integrity of a measurement instrument; multiple confirmatory factor analyses are used to assess measurement invariance across different groups, and to stamp the data as valid or invalid. The PAQLQ, a widely

Correspondence: Karen H. Sousa, RN, PhD, FAAN University of Colorado College of Nursing Aurora, CO 80045 Karen.Sousa@ucdenver.edu.

There are no conflicts of interest to declare.

used instrument having evidence to support reliability and validity was used separately in English- and Spanish-speaking groups. Traditional methods for evaluating measurement instruments have been less than thorough, and this paper demonstrates a well-developed approach allowing for confident comparisons between populations.

Keywords

patient outcomes; confirmatory factor analysis; measurement bias

Registered nurses and nurse researchers often use questionnaires to measure patient outcomes. When questionnaires or other multiple-item instruments have been developed using a relatively homogeneous sample, the suitability of even a psychometrically well-developed instrument for the new population comes into question (Corless, Nicholas, & Nokes, 2001; Staniszewska, Ahmed, & Jenkinson, 1999). This is problematic for any research involving comparisons among different groups. The questionnaire then has limited utility for providing interpretable findings both within the new population group (e.g., different ethnic, gender, or language groups) and for comparisons between the original and the new populations. The performance of the questionnaire in each subgroup must be considered. Introducing measurement invariance to validity discussions in nursing research can eliminate this limitation.

Two psychometric methods have been developed for establishing measurement invariance between different population groups: item response theory (Embretson & Reise, 2000) and confirmatory factor analysis (CFA; Meredith, 1993; Millsap, 2011; Widaman & Reise, 1997); the focus here is on methods of establishing measurement invariance using CFA. Although both approaches are useful, CFA is adapted more easily to instruments measuring multidimensional constructs; it can be applied also to questionnaires having binary, ordered category, or continuous response formats (Bollen, 1989; Muthén, 1984). Using available structural equation modeling software (e.g., AMOS, EQS, LISREL, Mplus) permits estimation of the multiple group CFA models needed to study measurement invariance.

Background

Juniper et al. (1996) developed the Paediatric Asthma Quality of Life Questionnaire (PAQLQ) as a conceptual model and measurement scale to measure quality of life in children living with asthma. They proposed that quality of life for a child living with asthma has three dimensions: symptoms, emotional function, and activity limitations. Decisions related to treatment, health resource allocations, and interventions are often based on PAQLQ-generated data. The PAQLQ has been translated into over 20 languages including Spanish. However, little is known about the measurement quality of the PAQLQ and the items as measured across groups because careful studies of the questionnaire's measurement invariance have not been undertaken. Juniper and colleagues at the Mapi Research Institute, Lyon, France (see www.qoltech.co.uk) used an extensive forward and backward translation procedure to develop the Spanish for Mexico version. They asserted that their version was equivalent to the original and that it could be used in studies of Spanish-speaking children. However, they did not suggest that it was appropriate to use the Spanish version for between-group comparisons with children answering the English version.

Many questionnaires, including the PAQOL, could be conceptually problematic when used to compare population groups (Johnson & Wolinsky, 1994; Markides et al., 1996; Mutran, Reed, & Sudha, 2001; Stewart & Napoles-Springer, 2000). Testing instruments for measurement invariance becomes a critical issue if outcomes are to be compared between groups, in this example English- and Spanish-speaking children living with asthma. Clearly,

confidence in the validity of the measurement instrument is critical when investing confidence in the fidelity of the between-group measure yielded by that instrument (Tann, Sousa, & Kwok, 2005). The PAQLQ influences the theory and practice of health care; however, the theory and practice are built upon an instrument whose validity has been tested insufficiently, resulting in potential measurement invariance.

Instruments must be designed to yield quantitative and replicable findings, specifically between cultural and language groups (Flaherty et al., 1988). Stewart and Napoles-Springer (2000) stated that a valid comparison of self-reported measures like the PAQLQ requires that the constructs have a similar meaning across groups. This semantic equivalence (identical meaning of each item after translation) is difficult to achieve and evaluate (Flaherty et al., 1988). The International Test Commission has emphasized that measurement equivalence cannot be assumed across language and cultural groups (Hambleton, Merenda, & Spielberger, 2005; Tanzer & Sim, 1999). “Test developers/ publishers should apply appropriate statistical techniques to (a) establish the equivalence of the different versions of the test or instrument, and (b) identify problematic components or aspects of the test or instrument which may be inadequate to one or more of the intended populations” (International Test Commission, 2010, guideline D7). Invariance analysis is used to address this issue.

Bias or lack of equivalence can be introduced into instruments through group differences in perceptions of the meaning of the measured items, differences in the constructs, or both. These differences in perception can result from group differences in the cognitive processes of answering items and the use of response scales, or from item differences resulting from inadequate translations (Hahn & Cella, 2003; Rogler, 1989). Observed lower levels of self-rated health in Latinos compared to other ethnic groups (Marin & Marin, 1991; Osmond, Vranizan, Schillinger, Stewart, & Bindman, 1996; Shetterly, Baxter, Mason, & Hamman, 1996) may be a function of a response pattern specific to Latinos (e.g., comparing one's health to that of others rather than viewing health as a personal state). Measures derived from instruments that do not take into account differences in processing between differing groups may not reflect real differences between those groups; this data-related uncertainty is antithetical to the translation of health care research into health care practice (Johnson et al., 1996).

In traditional two-group comparisons, there is a desire to draw conclusions as to how an intervention may affect each of the separate groups. Group comparisons of the mean scale scores of the measured items are justified to the extent that psychometric work indicates these comparisons approximate comparisons of means on the theoretical true score for the construct. Unambiguous interpretation of observed mean differences is dependent on the between-group equivalence of the underlying measurement model, and a comparison between different population groups--even with instruments that have been shown to have adequate reliability and validity within each group--requires making several traditionally untested assumptions described below (Flaherty et al., 1988; Hui & Triandis, 1985; Vandenberg & Lance, 2000).

If a measure is used to assess one thing in one group and something different in another group, comparing group means is like comparing apples to oranges (Chen, 2008). Differences in scores between groups *cannot* be used to infer group differences in the theoretical attribute unless the measured items accord with a particular set of invariance restrictions (Borsboom, 2006); if evidence supporting a measure's invariance is lacking, conclusions are at best ambiguous and at worst erroneous (Steenkamp & Baumgartner, 1998). Measurement invariance analysis provides a tool for developing psychometrically appropriate instruments for between-population-group comparisons.

With continuous variables, the most frequently used technique for testing measurement invariance is multiple-group CFA. The early statistical developments of this technique (e.g., Alwin & Jackson, 1981; Jöreskog, 1971) and the applications that followed were limited to the comparison of covariance structures; this work permitted researchers to establish that the theoretical constructs and the units in which they were measured were equivalent in the two populations, important for correlational studies. In later research (Meredith, 1993; Widaman & Reise, 1997), the technique was developed further, permitting the comparison of mean structures between the groups by establishing that the origin (0-point) of the theoretical constructs were equivalent in the two populations.

Meredith (1993), Widaman and Reise (1997), and Millsap (2011) developed a set of *hierarchical* tests to identify the extent to which measurement invariance can be established in the two populations. The most basic level of measurement invariance is *configural invariance* (Horn, McArdle, & Mason, 1983). The central requirement is that the same item must be an indicator of the same latent factor in each group; however, the magnitude of the factor loadings can differ across groups. When this level of invariance is achieved, roughly similar, but not identical, latent variables are present in the groups (Widaman & Reise, 1997), precluding clear comparisons between groups.

The second level of invariance is *factor loading invariance* (weak invariance). Factor loadings represent the strength of the linear relation between each factor and each of the associated items (Bollen, 1989; Jöreskog & Sörbom, 1999). When the loading of each item on the underlying factor is equal in two (or more) groups, the unit of measurement of the underlying factor is identical. This level of invariance does not require that the scales of the factors have a common origin. When *factor loading invariance* is met, relations between the factors can be compared across groups; one unit of change in one group is equal to one unit of change in another. However, the factor means of the scale still cannot be compared across groups, because the origin of the scale may differ.

The third level of invariance is *intercept invariance* (strong invariance). Intercepts represent the origin of the scale. In testing this form of invariance, each of the intercepts of the measured variables, in addition to factor loadings of the latent variables, is constrained to be equal across groups. This level of invariance is required for comparing latent means across groups (Millsap, 2011; Widaman & Reise, 1997). When this level of invariance is achieved, the scores from different groups have the same unit of measurement (factor loading) as well as the same origin (intercept). Without this level of invariance, it cannot be determined whether any difference between groups on factor means is a true group difference or a measurement artifact.

The fourth form of invariance is *residual invariance* (strict invariance). In testing this form of invariance, the residual (uniqueness or measurement error) associated with each measured variable, in addition to the factor loadings of the latent variables and the intercepts of measured variables, is constrained to be equal across groups. When this level of invariance holds, all group differences on the items are due *solely* to group differences on the common factors.

In summary, configural, factor loading, intercept, and residual invariance are the most commonly tested forms of measurement invariance. Modifications of the basic measurement invariance-testing procedure have been proposed to address instruments having more complex factor structures (e.g., second order factor models; Chen, West, & Sousa, 2006) or binary- or ordinal-response formats (e.g., Muthén & Christofferson, 1981).

Purpose

The goal of this study was to explain measurement invariance and illustrate how it can be tested using the English- and Spanish-language versions of the PAQLQ. These questionnaires have been used widely with pediatric asthma patients and have shown evidence of reliability and validity separately in both English and Spanish populations (Juniper et al., 1996; see www.qoltech.co.uk). Initially, the measurement structure of each of the scales was studied, and then analyses were conducted to establish the extent to which the same constructs are being assessed in the English- and Spanish-language forms of the instrument. Although different language groups were used, these procedures can be used to assess equivalence of constructs across age, gender, and ethnic groups.

Method

Participants

Since January 2000, the pulmonary division of Phoenix Children's Hospital has operated a mobile asthma clinic, the Phoenix Children's Hospital Breathmobile, for medically underserved, inner-city school age children. The program provides asthma case detection, education, and interventional treatment programs. The population consists of children 5-18 years old with asthma symptoms, identified through a case-detection process. A pediatric nurse practitioner, a respiratory therapist, an eligibility worker, and a bilingual pulmonary nurse staff the Breathmobile. All children referred are given the PAQLQ on their first visit. Depending on the primary language of the child, the PAQLQ is administered during the face-to-face interviews in either English or Spanish, following the protocol originally developed by Juniper (see www.qoltech.co.uk).

For this illustration medical records for all the children seen by the Breathmobile were abstracted. A sample of 607 primarily Mexican origin Hispanic children, with a completed PAQLQ, were used. The sample ranged in age from 6 to 18 years in age (M = 10.85, *SD* = 2.14) and was 51.4% male. Children from other ethnic backgrounds were excluded in the sample to permit valid group comparisons. Kirkman-Liff and Mondragón (1991) investigated language of interview in a large sample of Hispanics in the Southwest. They found that there were significant differences between Hispanics interviewed in English with those interviewed in Spanish on several health status items. Therefore, to demonstrate measurement invariance we were confident that our group distinction was appropriate. A total of 61.2% of the Hispanic children completed the questionnaire in Spanish, 38.8% in English.

Instrument

The English-language version of the PAQLQ and its Spanish- language version were made available to each of the children. Children were interviewed face-to-face in their preferred language based on the pediatric nurse practitioner's assessment. Norris et al. (1996) supported Marin and Marin's (1991) assumption that the language factor alone provides a valid and reliable indicator of acculturation. Both the English and Spanish versions of the scale have 23 items used to assess three domains: symptoms (10 items), emotional function (8 items), and activity limitation (5 items). Three of the 5 activity items potentially reflect different respondent-generated activities that are not consistent across children. For these items, each child is asked to nominate three important activities from a list of 35 potential activities (e.g., running, shopping) that are performed commonly, are important to the child, and are limiting because of the threat of exacerbating their asthma. The questions are: “How much have you been bothered by your asthma in [the activity] during the past week?” For ease of presenting this example, these items were excluded initially in the analysis because

of their unique structure and the necessity to make the strong assumption that the nature of the activity is comparable both within and across children (i.e., respondent 1's response to item 1 “running” is equivalent to respondent 2's response to item 1 “shopping”). Questions 5, 6, 7, 9, 11, 13, 15, 16, 17 18, 19, 20, 21, and 23 had a 7-point response format anchored by 1 = *all of the time* and 7 = *none of the time.* For questions 8, 10, 12, 14, and 22, the anchors were 1 = *extremely bothered* and 7 = *not bothered* (Table 1). The validity tests to which the PAQLQ has been subjected so far have shown it to have adequate internal consistency and predictive validity (Clarke et al., 1999; Juniper et al., 1996; Raat et al., 2005; Tauler et al., 2001).

Procedure and Statistical Analysis

Tests for the structure of the overall PAQLQ, for its measurement invariance between English- and Spanish-speaking children, and for the latent mean differences were based on the analysis of mean and covariance structures using multiple-group CFA. All CFAs were calculated using MPlus 5.2 (Múthen & Múthen, 2010).

Testing measurement invariance entails testing a series of hierarchically nested models; each pair of models in the sequence is nested because a set of parameters is constrained to be equal across groups, in the more restricted but not in the less restricted model. For example, in the configural invariance model, no constraints are placed on the *values* of the hypothesized factor loadings across groups, whereas the factor loading invariance model constrains the corresponding factor loadings to be equal in each group.

In assessing the adequacy of the fit of each model, both $\chi 2$ tests and goodness-of-fit indexes were used. The $\chi 2$ test assesses the magnitude of the discrepancy between the sample and fitted covariance matrices and the sample and fitted mean vectors. A significant test result indicates that the fit is poor and that the model may not be appropriate for the data. However, moderate discrepancies from normality in the data also lead to rejection of the model using the $\chi 2$ test (West, Finch, & Curran, 1995). Furthermore, when the sample size is large, a small discrepancy that may be of no practical or theoretical interest can lead to similar rejection. Consequently, the $\chi 2$ test of fit was supplemented with three fit indexes that showed good performance in a simulation study by Hu and Bentler (1998). The root mean squared error of approximation (RMSEA; Steiger, 1990) is a measure of the estimated discrepancy between the population and model implied population covariance matrices per degree of freedom. Browne and Cudeck (1993) suggested that values of the RMSEA of .05 or less indicate a close fit and .08 or less indicate adequate fit. The standardized root mean square residual (SRMR; Hu & Bentler, 1998) is a measure of the average of the standardized fitted residuals. It ranges from 0 to 1.00, and a value of less than .08 is conventionally taken as an indication of adequate fit. The comparative fit index (CFI; Bentler, 1990) ranges from 0 (poor fit) to 1.00 (perfect fit) and is derived from a comparison of a restricted model in which restrictions are imposed on the data with a baseline model in which all pairs of observed variables are assumed to be mutually uncorrelated. Hu and Bentler (1999) suggested the use of .95 as a criterion for adequate fit.

To compare the fit for the two nested models representing different levels in the hierarchy of measurement invariance, the likelihood ratio test (the $\chi 2$ difference test; Bentler & Bonett, 1980) was used. If significant, the likelihood ratio test (the difference between the $\chi 2$ statistics associated with the more and less restricted models) suggests that the constraints on the more restricted model may be too strict. Large sample sizes can lead to the detection of significant differences between two models of a tiny magnitude that is of no practical importance. Based on a simulation study, Cheung and Rensvold (2002) concluded that a difference greater than .01 in the CFI would indicate a meaningful change in model fit.

Latent mean difference test—Once measurement invariance is established, multigroup CFA may be used to test whether the latent factor means differ across the groups. To test the latent construct mean differences, a combined mean and covariance structure model must be used (Bentler, 1989; Bollen, 1989; Sörbom, 1978). To estimate the difference between the factor means, one group is chosen as a reference or baseline group and the latent means are set to zero. The latent means of the other group, which represent the difference between the factor means in the two groups, are estimated. The significance test (Wald or *z-test)* for the latent means of the second group provides a test for significance of the difference between the means of the two groups on the latent construct (Aiken, Stein, & Bentler, 1994).

Qualitative itemmetric analyses—Since previous work on the structure of the PAQLQ had not been reported, separate itemmetric analysis were run (Angleitner, John, & Löhr, 1986) of the English- and Spanish-language versions of the PAQLQ. The checks on the Spanish version were conducted by bilingual research assistants from Mexico who represented the local dialects of Spanish in the Southwest United States, where the study took place.

For the English-language version, each of the items was hypothesized to measure a different facet of asthma quality of life: symptoms, emotional function, or activity limitations (see Juniper et al., 1996 for list of items and domain assessed). Each pair of items was examined for similarity of wording, identifying items that were potentially more similar than would be expected based on their assessment of a single construct. Two pairs of items (16 and 20, and 15 and 19; Table 2) were identified as having potentially high correlated residuals due to their almost identical wording and content.

Two bilingual Mexican-origin Spanish speakers examined each pair of items in the Spanish version for similarity of wording and identified potentially problematic items (see www.qoltech.co.uk) for wording of items). Again, pairs 16 and 20, and 15 and 19 were identified as possibly having correlated residuals due to their almost identical wording and content.

Quantitative analyses—Items that were common across respondents were analyzed, and the three participant-nominated items (items 1-3) were deleted. Several models were tested to assess and establish the overall factor structure of the scale separately in English and Spanish. Also estimated was a two-factor model in which items 1-3 were allowed to load on a separate factor to determine whether the 3 self-nominated items (1-3) were related to the 20 items (4-23) that were consistent across respondents. The correlations between the two factors were .292 in the English-speaking group and .147 in the Spanish-speaking group. These results support the decision to discard the three self-nominated items as they were measuring a different factor than the remaining items.

Following Juniper et al. (1996), the scale was hypothesized to consist of three related domains that would be represented by a structure with three correlated factors (Figure 1). The hypothesized factor model was specified as follows: (a) each item would have a nonzero loading on the factor (symptoms, emotional function, and activity limitations) that it was designed to measure and a zero loading on each of the other factors; and (b) error terms associated with each item would be uncorrelated. Identification of the model is required for estimation in CFA. A model is identified if there is a unique numerical solution for each of the parameters (Ullman, 2001). In the context of multiple group CFA, the marker variable strategy is used to identify the scale of the measurement models. One of the factor loadings (marker variable) is set to a value of 1 for each factor. Since there was no theoretical basis for choosing a marker, the first indicator of each construct was used as the marker.

Results

Both the English and Spanish PAQLQ questionnaires were hypothesized to assess three domains that are represented as three correlated factors. The model depicted in Figure 2 in which each of the items was expected to load on one factor corresponding to the domain was used as specified in Juniper et al. (1996, p. 46). Each pair of the three factors was allowed to correlate. To preserve the possibility of having different factor structures in the English and Spanish groups, the configural model was tested (Horn & McArdle, 1992) in which the factor loadings, correlations between factors, and unique variances could differ between the English and Spanish speaking groups. The factor loadings for English- and Spanish-speaking children are found in Table 3. The overall fit of the three-factor model was not acceptable based on the $\chi 2$ and fit indices ($\chi 2(334) = 848.89$, $p < .001$; CFI = .92; RMSEA = .07; SRMR = .04). Further, the solution was improper: There were correlations greater than 1.0 between the latent factors activity limitation and emotional function, and activity limitation and symptoms in both the English- and Spanish-speaking groups. These results suggested that a single general quality-of-life factor may be sufficient to account for the data. No additional reliable information appears to be represented by the three specific domains.

To probe this observation, the fit of the overall quality-of-life single factor model was tested. Included were items 4-23 for both the English and Spanish versions. The overall fit of the model was not adequate in terms of the chi-square and CFI indices ($\chi 2(340) = 912.05$, $p < .001$; CFI = .91; RMSEA = .07; SRMR = .05). The factor loadings for the single-factor model of the English and Spanish versions are presented in Table 2.

Observations from the itemmetric analyses were used to develop highly restricted hypotheses that might lead to improvements in the factor structure, proceeding one step at a time in model modification (MacCallum, 1986). The single factor model was tested first, allowing the residuals of items 16 and 20 to correlate in both English and Spanish. The fit of the model improved ($\chi 2(338) = 727.12$, $p < .001$; CFI = .94; RMSEA = .06; SRMR = .04) and was significantly better than the fit of the original single factor model ($\Delta\chi 2(2) = 184.93$, $p < .001$). The correlated residuals were significant for both English and Spanish (r = .33, $p < .001$; $r = .30$, $p < .001$, respectively). Next, residuals of items 15 and 19 were allowed to correlate for both the English and Spanish groups. The fit of the model again improved ($\chi 2(336) = 593.19$, $p < .001$; CFI = .96; RMSEA = .05; SRMR = .04) and was significantly better than the fit of the previous model ($\Delta\chi 2(2) = 133.93$, $p < .001$). The correlated residuals were significant for both English and Spanish (r = .17, $p < .001$, $r = .28$, $p < .001$, respectively). Item 15 represented the emotional functioning domain, whereas item 19 represented the activity limitation domain. Since these items loaded on different factors, they represent a likely source of some of the problems in estimating the earlier models noted above.

Structural Invariance of the One-Factor Model

The structural invariance of the final one factor model from the CFA was tested with the correlated residuals of items 16 and 20, and 15 and 19 for both the English and Spanish scales. The factor loading of item 4 was set to 1.0 and the intercept for item 4 was set equal across the two language groups to identify the model. Step 1 began with the baseline configural model, which allows all the other factor loadings and intercepts, as well as the measurement error variances and correlated measurement errors, to be estimated freely. The pattern of fixed and free factor loadings was constrained to be the same across groups, but different estimates were allowed for the corresponding parameters in the different groups. The fit of this model was acceptable ($\chi 2(336) = 593.19$, $p < .001$; CFI = .96; RMSEA = .05; SRMR = .04).

In step 2, the factor loadings were constrained to be equal across the English and Spanish scales. This level of invariance was nested within Model 1. The fit of this model was also acceptable ($\chi 2(355) = 604.53$, $p < .001$; CFI = .96; RMSEA = .05; SRMR = .04) and did not differ from that of the baseline configural model ($\Delta\chi 2(19) = 11.34$, n.s.). This result indicates that a 1-unit change on the latent quality-of-life factor is not associated with differential change on any corresponding items on the English and Spanish questionnaires. The factor loadings were invariant across the two groups.

In step 3, the intercepts were constrained to be equal across groups. This condition is required to detect potential differences in the intercepts of the measured variable between the groups. The fit of this model was also acceptable ($\chi 2(374) = 621.47$, $p < .001$; CFI = .96; RMSEA = .05; SRMR = .04) and did not differ from that of the previous model ($\Delta\chi 2(19) = 16.94$, n.s.). The result indicates that the zero-points on the corresponding latent factor for the English and Spanish items did not differ appreciably and that comparisons of means may be conducted.

In step 4, the correlated residuals were constrained to be equal across groups. The fit of this model was still acceptable ($\chi 2(376) = 627.14$, $p < .001$; CFI = .96; RMSEA = .05; SRMR = .04) and did not differ significantly from that of the model in step 3 ($\Delta\chi 2(2) = 5.67$, n.s.).

In step 5, the residual variances were constrained to be equal across groups. The fit of this model was acceptable ($\chi 2(396) = 638.83$, $p < .001$; CFI = .96; RMSEA = .05; SRMR = .05) and did not differ significantly from that of the previous model ($\Delta\chi 2(20) = 11.69$, n.s.). These results indicate that measurement invariance was obtained for the one-factor model with two correlated residuals across the English- and Spanish-language versions of the PAQLQ.

Equality of Factor Means

After establishing invariance between the English- and Spanish-language groups, the equality of latent factor means of the two groups was tested. In the preceding model, the latent factor mean for the English group was set to 0, and the latent factor mean for the Spanish group was free to vary. The estimated factor mean for the Spanish group is the mean difference between the latent factors for English and Spanish. The estimated value was .185, $p = .054$. Thus, while the Spanish-speaking group had a higher factor mean than the English-speaking group, this difference was only marginally statistically significant.

Discussion

Confirmatory factor analysis is used to evaluate the structural integrity of a measurement instrument; multigroup CFA is used to assess measurement invariance across different age, cultural, ethnic, or language groups. The PAQLQ, a widely used instrument with evidence supporting its reliability and validity separately in English- and Spanish-language groups, was used. The hypothesized measurement structure of the PAQLQ had not been tested systematically previously. There were problems with the hypothesized three-factor structure; a one-factor structure with the same two correlated residuals in both language groups to account for wording similarity provided an adequate fit in each language. The test of measurement invariance for the one-factor model showed that the English and Spanish versions of the scale met the criteria for measurement invariance. The level of strict invariance (equal factor loadings, intercepts, and residual variances between groups) was achieved, which indicates that any differences between the two language groups are accounted for by group differences on the common factor. This finding is encouraging, indicating that mean differences in the quality of life construct assessed by the English- and Spanish-language versions of the PAQLQ may be compared directly. The slightly higher

mean level in the Spanish-speaking group relative to the English-speaking group can be interpreted as indicating that the less acculturated Spanish-speaking group perceived a higher quality of life. However, this difference was only marginally significant.

Although a strict measurement invariance was established with the PAQLQ with respect to test language, an outcome of strict measurement invariance will often not be the case. When a test of a level of measurement invariance fails, it is only appropriate to interpret the meaning of the scale at the previous level in the hierarchy at which measurement invariance was established.

For example, if the test of the equality of the intercepts failed (step 3), then the measurement scales have only demonstrated that scales could be interpreted at the level of weak invariance (step 2). Given such an outcome, the difference between slopes representing relationships between constructs within each group could be compared, for example, using moderated multiple regression (Cohen, Cohen, West, & Aiken, 2003). Alternatively, Yoon and Millsap (2007) have developed a specification search procedure to identify the noninvariant items when only partial measurement invariance is achieved. Simulation studies have shown that this procedure works well if the number of invariant items is small, and group differences and the sample size in each group are large.

In the present study, only the most prominent potential source of invariance was investigated: test language. Space limitations precluded the investigation of invariance across age groups (e.g., young children vs. teenagers) or gender. It is particularly important to investigate potential group differences if there is theory or empirical work to suggest that the constructs may differ in the subgroups and the goal of the research is to compare between group differences. Other limitations are that no formal measurement was used to assess acculturation in the children and the data used for this illustration were collected for routine screening of the children during a visit to the Breathmobile, not for research.

In the absence of measurement invariance procedures, comparisons between different age, cultural, ethnic, gender, language, or other groups are difficult to make. Vandenberg and Lance (2000) note comparisons between groups still require several assumptions regarding equivalence. These assumptions are examined rarely and, if violated, can render interpretations of between-population-group comparisons on the nonequivalent measures highly suspect (Bollen, 1989; Vandenberg & Lance, 2000). Measurement invariance analysis provides tests of each of these assumptions. Traditional methods for evaluating measurement instruments have been less thorough, and a well-developed approach was presented here to address these concerns. Testing for and assuring measurement invariance is essential for a profession such as nursing, where group comparisons are frequently of interest.

Acknowledgments

This study was supported partially by funds from the National Institute of Nursing Research (NIH #1R15NR010632-010).

References

Aiken LS, Stein JA, Bentler PM. Structural equation analyses of clinical subpopulation differences and comparative treatment outcomes: Characterizing the daily lives of drug addicts. Journal of Consulting and Clinical Psychology. 1994; 62:488–499. [PubMed: 8063975]

Alwin, DF.; Jackson, DJ. Applications of simultaneous factor analysis to issues of factorial invariance.. In: Jackson, D.; Borgotta, E., editors. Factor analysis and measurement in sociological research: A multi-dimensional perspective. Sage; Beverly Hills, CA: 1981. p. 68-119.

Angleitner, A.; John, OP.; Löhr, FJ. It's how you ask and what you ask: An itemmetric analysis of personality questionnaires.. In: Angleitner, A.; Wiggins, JS., editors. Personality assessment via questionnaires. Springer; New York, NY: 1986. p. 61-108.

Bentler, PM. EQS: Structural equations program manual. BMDP Statistical Software; Los Angeles, CA: 1989.

Bentler PM. Comparative fit indexes in structural models. Psychological Bulletin. 1990; 107:238–246. [PubMed: 2320703]

Bentler PM, Bonett DG. Significance tests and goodness of fit in the analysis of covariance structures. Psychological Bulletin. 1980; 88:588–606.

Bollen, KA. Structural equations with latent variables. Wiley; New York, NY: 1989.

Borsboom D. When does measurement invariance matter? Medical Care. 2006; 44:S176–S181. [PubMed: 17060825]

Browne, MW.; Cudeck, R. Alternative ways of assessing model fit.. In: Bollen, KA.; Long, JS., editors. Testing structural equation models. Sage; Newbury Park, CA: 1993. p. 136-162.

Chen FF. What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. Journal of Personality and Social Psychology. 2008; 95:1005–1018. [PubMed: 18954190]

Chen FF, West SG, Sousa KH. A comparison of bifactor and second-order models of quality of life. Multivariate Behavioral Research. 2006; 41:189–224.

Cheung GW, Rensvold RB. Evaluating goodness-of-fit indexes for testing measurement invariance. Structural Equation Modeling. 2002; 9:233–255.

Clarke E, Sulaiman S, Chew Fook T, Shek Lynette Pei C, Mital R, Lee BW. Pediatric asthma quality of life questionnaire: Validation in children from Singapore. Asian Pacific Journal of Allergy and Immunology. 1999; 17:155–161. [PubMed: 10697253]

Cohen, J.; Cohen, P.; West, SG.; Aiken, LS. Applied multiple regression/correlation analysis for the behavioral sciences. 3rd ed.. Erlbaum; Mahwah, NJ: 2003.

Corless IB, Nicholas PK, Nokes KM. Issues in cross-cultural quality of life research. Journal of Nursing Scholarship. 2001; 33:15–20. [PubMed: 11253575]

Embretson, SE.; Reise, SP. Item response theory for psychologists. Erlbaum; Mahwah, NJ: 2000.

Flaherty JA, Gaviria FM, Pathak D, Mitchell T, Wintrob R, Richman JA, Birz S. Developing instruments for cross-cultural psychiatric research. The Journal of Nervous and Mental Disease. 1988; 176:257–263. [PubMed: 3367140]

Hahn MA, Cella D. Health outcomes assessment in vulnerable populations: measurement challenges and recommendations. Archives of Physical Medicine and Rehabilitation. 2003; 84:S35–S42. [PubMed: 12692770]

Hambleton, RK.; Merenda, P.; Spielberger, C., editors. Adapting educational and psychological tests for cross-cultural assessment. Erlbaum; Hillsdale, NJ: 2005.

Horn JL, McArdle JJ. A practical and theoretical guide to measurement invariance in aging research. Experimental Aging Research. 1992; 18:117–144. [PubMed: 1459160]

Horn JL, McArdle JJ, Mason R. When is invariance not invariant: A practical scientist's look at the ethereal concept of factor invariance. Southern Psychologist. 1983; 4:179–188.

Hu LT, Bentler PM. Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. Psychological Methods. 1998; 3:424–453.

Hu LT, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling. 1999; 6:1–55.

Hui CH, Triandis HC. Measurement in cross-cultural psychology: A review and comparison of strategies. Journal of Cross-Cultural Psychology. 1985; 16:131–152.

International Test Commission. International Test Commission guidelines for translating and adapting tests. 2010. http://www.intestcom.org/upload/sitefiles/40.pdf

Johnson RJ, Wolinsky FD. Gender, race, and health: The structure of health status among older adults. The Gerontologist. 1994; 34:24–35. [PubMed: 8150305]

Johnson, TP.; O'Rourke, D.; Chavez, N.; Sudman, S.; Warnecke, RB.; Lacy, L.; Horm, J. Cultural variations in the interpretation of health survey questions.. In: Warnecke, R., editor. Health survey

research methods: Conference proceedings. National Center for Health Statistics; Hyattsville, MD: 1996. p. 57-62.

Jöreskog KG. Simultaneous factor analysis in several populations. Psychometrika. 1971; 36:409–426.

Jöreskog, KG.; Sörbom, D. LISREL 8: User's reference guide. 2nd ed.. Scientific Software International, Inc.; Chicago, IL: 1999.

Juniper EF, Guyatt GH, Feeny DH, Ferrie PJ, Griffith LE, Townsend M. Measuring quality of life in children with asthma. Quality of Life Research. 1996; 5:35–46. [PubMed: 8901365]

Kirkman-Liff B, Mondragón D. Language of interview: Relevance for research of Southwest Hispanics. American Journal of Public Health. 1991; 81:1399–1404. [PubMed: 1951794]

MacCallum RC. Specification searches in covariance structure modeling. Psychological Bulletin. 1986; 100:107–120.

Marin, G.; Marin, BV. Research with Hispanic populations. Sage; Newbury Park, CA: 1991.

Marin G, Sabogal F, Marin BV, Otero-Sabogal R, Perez-Stable EJ. Development of a short acculturation scale for Hispanics. Hispanic Journal of Behavioral Sciences. 1987; 9:183–205.

Markides KS, Stroup-Benham CA, Goodwin JS, Perkowski LC, Lichtenstein M, Ray LA. The effect of medical conditions on the functional limitations of Mexican-American elderly. Annals of Epidemiology. 1996; 6:386–391. [PubMed: 8915469]

Meredith W. Measurement invariance, factor analysis and factorial invariance. Psychometrika. 1993; 58:525–543.

Millsap, RE. Statistical approaches to measurement invariance. Routledge; New York, NY: 2011.

Múthen B. A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. Psychometrika. 1984; 49:115–132.

Múthen B, Christofferson A. Simultaneous factor analysis of dichotmous variables in several groups. Psychometrika. 1981; 46:407–419.

Múthen, LK.; Múthen, BO. Mplus user's guide. Author; Los Angeles, CA: 2010.

Mutran EJ, Reed PS, Sudha S. Social support: Clarifying the construct with applications for minority populations. Journal of Mental Health Aging. 2001; 7:137–164.

Norris AE, Ford K, Bova CA. Psychometrics of a brief acculturation scale for Hispanics in a probability sample of urban Hispanic adolescents and young adults. Hispanic Journal of Behavioral Sciences. 1996; 18:29–38.

Osmond DH, Vranizan K, Schillinger D, Stewart AL, Bindman AB. Measuring the need for medical care in an ethnically diverse population. Health Service Research. 1996; 31:551–571.

Raat H, Bueving HJ, de Jongste JC, Grol MH, Juniper EF, van der Wouden JC. Responsiveness, longitudinal- and cross-sectional construct validity of the Pediatric Asthma Quality of Life Questionnaire (PAQLQ) in Dutch children with asthma. Quality of Life Research. 2005; 14:265–272. [PubMed: 15789960]

Rogler LH. The meaning of culturally sensitive research in mental health. American Journal of Psychiatry. 1989; 146:296–303. [PubMed: 2919686]

Shetterly SM, Baxter J, Mason LD, Hamman RF. Self-rated health among Hispanic vs non-Hispanic white adults: The San Luis Valley Health and Aging Study. American Journal of Public Health. 1996; 86:1798–1801. [PubMed: 9003141]

Sörbom, D. A general method for studying differences in factor means and factor structure between groups.. In: Joereskog, KG.; Soerbom, D., editors. Advances in factor analysis and structural equation modeling. Abt Books; Cambridge, MA: 1978. p. 207-218.

Staniszewska S, Ahmed L, Jenkinson C. The conceptual validity and appropriateness of using health-related quality of life measures with minority ethnic groups. Ethnicity and Health. 1999; 4:51–63. [PubMed: 10887462]

Steenkamp JEM, Baumgartner H. Assessing measurement invariance in cross-national consumer research. Journal of Consumer Research. 1998; 25:78–90.

Steiger JH. Structural model evaluation and modification: An interval estimation approach. Multivariate Behavioral Research. 1990; 25:173–180.

Stewart AL, Napoles-Springer A. Health-related quality-of-life in diverse population groups in the United States. Medical Care. 2000; 38:102–124.

Tann S, Sousa KH, Kwok O. Cross-cultural model testing: Quality of life in a Latino population living with HIV. Hispanic Health Care International. 2005; 3:103–115.

Tanzer NK, Sim COE. Adapting instruments for use in multiple languages and cultures: A review of the ITC guidelines for test adaptations. European Journal of Psychological Assessment. 1999; 15:258–269.

Tauler E, Vilagut G, Grau G, Gonzalez A, Sanchez E, Figueras G, Alonso J. The Spanish version of the Paediatric Asthma Quality of Life Questionnaire (PAQLQ): Metric characteristics and equivalence with the original version. Quality of Life Research. 2001; 10:81–91. [PubMed: 11508478]

Ullman, JB. Structural equation modeling.. In: Tabachnick, BG.; Fidell, LS., editors. Unsing multivariate statistics. Allyn and Bacon; Boston, MA: 2001. p. 653-771.

Vandenberg RJ, Lance CE. A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. Organizational Research Methods. 2000; 3:4–69.

West, SG.; Finch, JF.; Curran, PJ. Structural equation models with nonnormal variables: Problems and remedies.. In: Hoyle, RH., editor. Structural equation modeling: Concepts, issues, and applications. Sage; Thousand Oaks, CA: 1995. p. 56-75.

Widaman, KF.; Reise, SP. Exploring the measurement invariance of psychological instruments: Applications in the substance use domain.. In: Bryant, KJ.; Windle, M.; West, SG., editors. The science of prevention: Methodological advances from alcohol and substance abuse research. American Psychological Association; Washington, DC: 1997. p. 281-324.

Yoon M, Millsap RE. Detecting violations of factorial invariance using data-based specification searches: A Monte Carlo study. Structural Equation Modeling. 2007; 14:435–463.

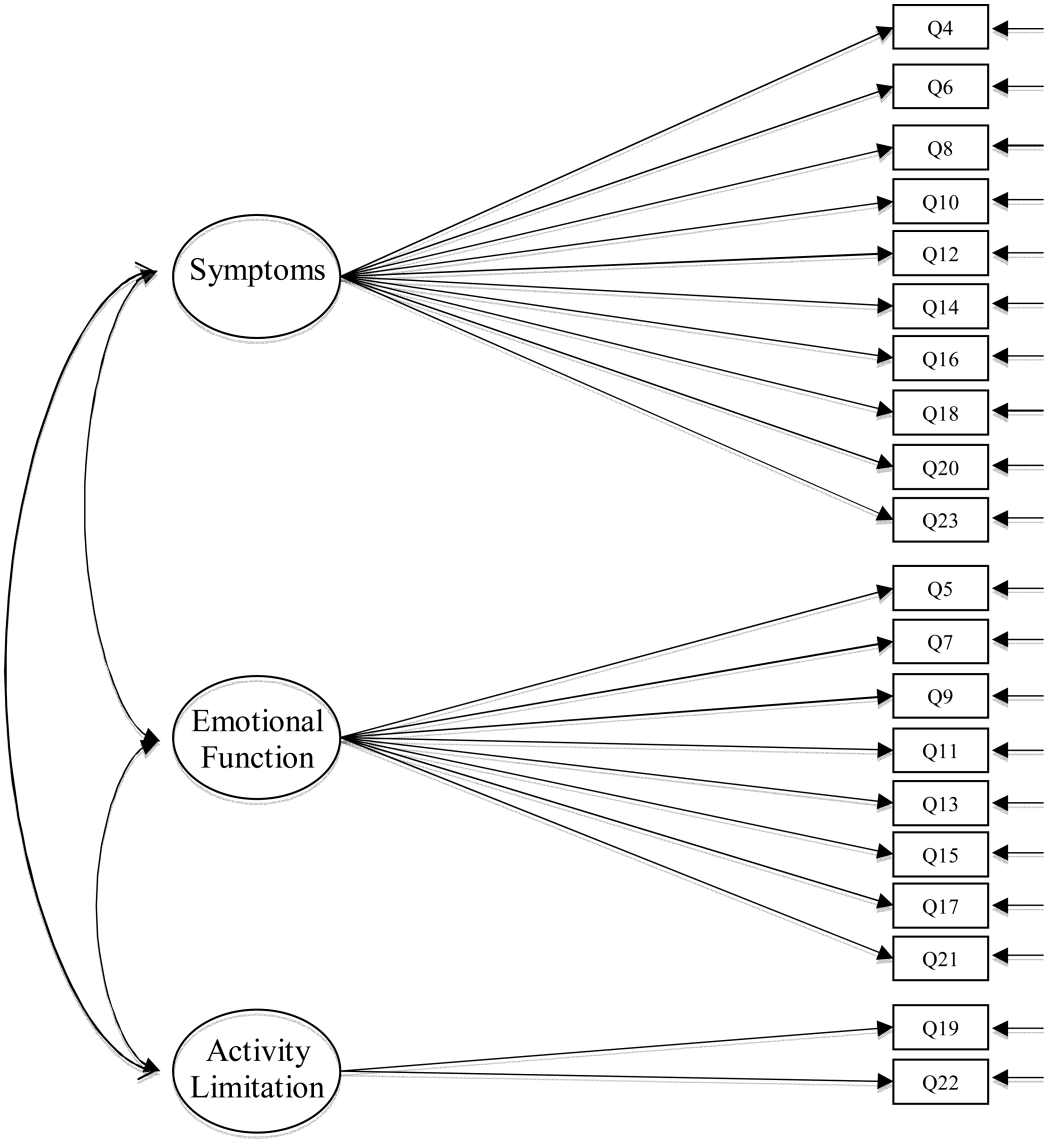

Figure 1.
Three-Factor Hypothesized Model

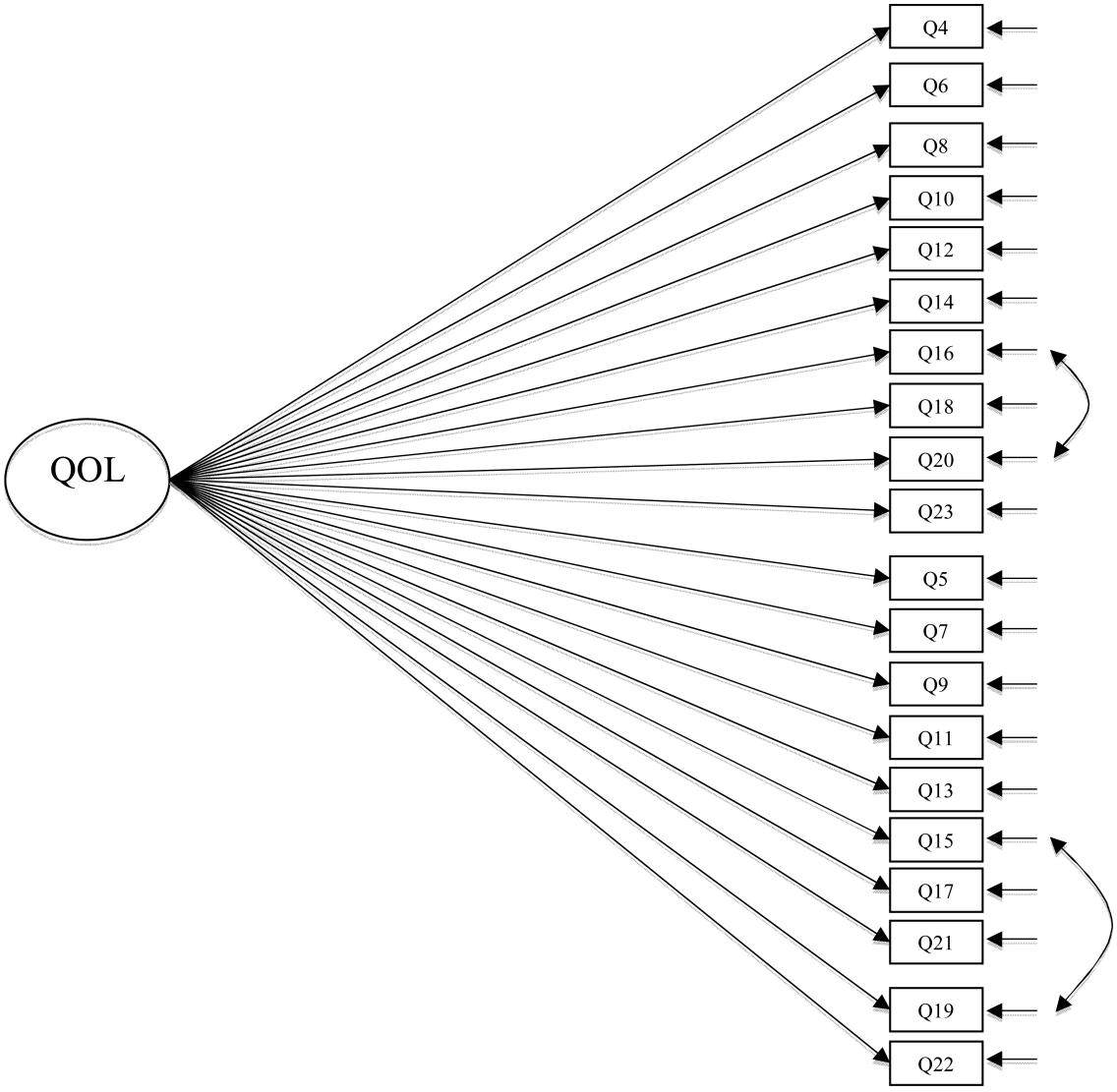

Figure 2.
Single-Factor Model

Table 1

Scale Items and Descriptive Statistics for English- (n = 234) and Spanish-Speaking (n = 373) Participants

Item	Mean	SD	Skew	Kurtosis
1. Bothered	3.50 (3.52)	2.04 (2.00)	.10 (-.03)	-1.06 (-.93)
2. Bothered	3.33 (3.27)	2.05 (2.04)	.003 (-.04)	-.95 (-.95)
3. Bothered	3.08 (3.06)	2.19 (2.17)	.09 (-.01)	-1.14 (-1.16)
4. Coughing	4.12 (4.31)	2.13 (2.07)	-.03 (-.18)	-1.44 (-1.34)
5. Frustrated	3.81 (3.93)	2.09 (2.02)	.27 (.13)	-1.34 (-1.30)
6. Tired	3.92 (4.11)	1.96 (1.86)	.13 (-.04)	-1.18 (-1.02)
7. Worried, Concerned or Troubled	4.36 (4.43)	2.11 (2.01)	-.22 (-.25)	-1.33 (-1.22)
8. Asthma attacks	4.96 (5.13)	2.11 (2.12)	-.22 (-.76)	-1.33 (-.89)
9. Angry	4.65 (4.66)	2.15 (2.11)	-.31 (-.41)	-1.39 (-1.19)
10. Wheezing	4.43 (4.72)	2.12 (2.07)	-.15 (-.34)	-1.42 (-1.28)
11. Irritable	4.60 (4.84)	2.07 (1.90)	-.32 (-.44)	-1.27 (-1.02)
12. Tightness in your chest	4.22 (4.52)	2.07 (2.02)	-.02 (-.29)	-1.31 (-1.21)
13. Different or left out	4.93 (5.16)	2.17 (2.09)	-.52 (-.77)	-1.25 (-.82)
14. Shortness of breath	4.43 (4.52)	1.94 (1.91)	-.22 (-.21)	-1.17 (-1.11)
15. Frustrated	4.12 (4.44)	2.26 (2.19)	-.06 (-.30)	-1.52 (-1.38)
16. Wake you up during the night	4.74 (5.07)	2.18 (2.13)	-.30 (-.67)	-1.45 (-1.00)
17. Uncomfortable	4.38 (4.55)	2.10 (1.98)	-.20 (-.30)	-1.29 (-1.20)
18. Out of breath	4.20 (4.64)	2.02 (1.88)	.02 (-.40)	-1.36 (-.97)
19. Couldn't keep up	4.17 (4.54)	2.21 (2.13)	-.04 (-.36)	-1.46 (-1.27)
20. Trouble sleeping at night	4.56 (5.06)	2.31 (2.01)	-.26 (-.64)	-1.54 (-.96)
21. Frightened	4.61 (4.99)	2.39 (2.20)	-.34 (-.63)	-1.54 (-1.14)
22. Bothered by your asthma	3.91 (4.22)	2.03 (1.96)	.13 (-.16)	-1.36 (-1.24)
23. Deep breath	4.42 (4.58)	2.16 (2.03)	-.17 (-.34)	-1.42 (-1.21)

Notes. Descriptive statistics for the English-speaking participants are presented next to the corresponding item. Descriptive statistics for the Spanish-speaking participants are presented in parentheses.

Table 2

Standardized Factor Loadings for Three-Factor Model Excluding Questions 1-3

	English			Spanish		
Item	**Symptom**	**Emotion**	**Activity**	**Symptom**	**Emotion**	**Activity**
4	.54			.52		
6	.62			.57		
8	.66			.58		
10	.64			.65		
12	.72			.74		
14	.70			.68		
16	.66			.66		
18	.68			.71		
20	.70			.68		
23	.72			.69		
5		.76			.74	
7		.71			.62	
9		.64			.66	
11		.68			.66	
13		.71			.62	
15		.75			.71	
17		.72			.74	
21		.71			.61	
19			.77			.70
22			.73			.67

Nurs Res. Author manuscript; available in PMC 2013 May 01.

Table 3

Standardized Factor Loadings for the Single-Factor Model Excluding Questions 1-3

Items	English	Spanish
4	.54	.50
5	.76	.74
6	.63	.57
7	.71	.61
8	.71	.61
9	.63	.64
10	.62	.64
11	.68	.65
12	.71	.72
13	.70	.60
14	.70	.66
15	.74	.68
16	.65	.64
17	.72	.74
18	.68	.69
19	.77	.70
20	.69	.66
21	.71	.61
22	.73	.72
23	.71	.67

Table 4

Summary of Fit Statistics for Testing Measurement Invariance of Second-Order Factor Model of Quality of Life

Model (*N* = 607) (234 vs. 373)	χ^2 (df)	RMSEA	SRMR	CFI	Model Comparison	$\Delta\chi^2$ (Δdf)	*p*
Model 1 Configural Invariance	593.19 (336)	.05	.04	.96	-	-	-
Model 2 Factor Loadings Invariant	604.53 (355)	.05	.04	.96	2 vs. 1	11.34 (19)	0.91
Model 3 Factor Loadings & Intercepts Invariant	621.47 (374)	.05	.04	.96	3 vs. 2	16.94 (19)	0.59
Model 4 Factor Loadings, Intercepts, & correlated residuals Invariant	627.14 (376)	.05	.04	.96	4 vs. 3	5.67 (2)	0.06
Model 5 Factor Loadings, Intercepts, correlated residuals, & residual variances Invariant	638.83 (396)	.05	.05	.96	5 vs. 4	11.69 (20)	0.93

Notes. RMSEA = root mean square error of approximation, SRMR = standardized root mean square residual, CFI = comparative fit index, $\Delta\chi^2$ = likelihood ratio test (chi square difference test)