Published in final edited form as:

Appl Res Qual Life. 2017 June ; 12(2): 251–288. doi:10.1007/s11482-017-9516-9.

Methodological issues in measuring subjective well-being and quality-of-life: Applications to assessment of affect in older, chronically and cognitively impaired, ethnically diverse groups using the Feeling Tone Questionnaire

Jeanne A. Teresi[1,2,3], **Katja Ocepek-Welikson**[2], **John A. Toner**[1], **Marjorie Kleinman**[4], **Mildred Ramirez**[2,3], **Joseph P. Eimicke**[2,3], **Barry J. Gurland**[1], and **Albert Siu**[5]

[1]Columbia University Stroud Center at New York State Psychiatric Institute

[2]Research Division, Hebrew Home at Riverdale; RiverSpring Health

[3]Weill Cornell Medical Center, Department of Geriatrics and Palliative Medicine

[4]New York State Psychiatric Institute, Division of Child and Adolescent Psychiatry

[5]Brookdale Department of Geriatrics and Palliative Medicine, Mount Sinai School of Medicine

Abstract

Quality of life assessment includes measurement of positive affect. Methods artifacts associated with positively and negatively worded items can manifest as negative items loading on a second factor, despite the conceptual view that the items are measuring one underlying latent construct. Negatively worded items may elicit biased responses. Additionally, item-level response bias across ethnically diverse groups may compromise group comparisons. The aim was to illustrate methodological approaches to examining method factors and measurement equivalence in an affect measure with 9 positively and 7 negatively worded items: The Feeling Tone Questionnaire (FTQ). The sample included 4,960 non-Hispanic White, 1,144 non-Hispanic Black, and 517 Hispanic community and institutional residents receiving long-term supportive services. The mean age was 82 (s.d.=11.0); 73% were female. Two thirds were cognitively impaired. Methods effects were assessed using confirmatory factor analyses (CFA), and reliability with McDonald's omega and item response theory (IRT) generated estimates. Measurement equivalence was examined using IRT-based Wald tests. Methods effects associated with negatively worded items were observed; these provided little IRT information, and as a composite evidenced lower reliability. Both 13 and 9 item positive affect scales performed well in terms of model fit, reliability, IRT information, and evidenced little differential item functioning of high magnitude or impact. Both CFA and IRT approaches provided complementary methodological information about scale performance. The 9-item affect scale based on the FTQ can be recommended as a brief quality-of-

Correspondence concerning this article should be addressed to: Jeanne A. Teresi, Ed.D., Ph.D., Columbia University Stroud Center at New York State Psychiatric Institute, 1051 Riverside Drive, Box 42, Room 2714, New York, New York, 10032-3702; Teresimeas@aol.com; jat61@columbia.edu, Phone: 718-581-1132; Fax: 718-543-2477.

This paper was based on a presentation at a preconference: Subjective well-being assessment in minority aging research, delivered on November 18, 2015 at the Gerontological Society of America meetings in Orlando, sponsored by the NIA Resource Centers for Minority Aging Research Coordinating Center (2P30AG021684).

Conflict of Interest: The authors declare that they have no conflicts of interest.

life measure among frail and cognitively impaired individuals in palliative and long-term care settings.

Keywords

methods; methods factors; item response theory; differential item functioning; ethnicity; quality-of-life; affect; Feeling Tone Questionnaire

INTRODUCTION

The measurement of quality-of-life in older persons has been defined and studied in several contexts (Albert and Teresi 2002; Brod et al. 1999; Gurland et al. 2014; Hickey et al. 2005; Lawton 1983; Lawton et al. 1999). One model of quality of life proposes that biological and psychological systems, influenced by social and environmental forces operate through health related and health independent pathways to impact quality of life (Gurland and Gurland 2009a; Gurland and Gurland 2009b). The complexity of quality of life assessment is illustrated by the differential impact of domains that qualitatively express distress and of those indicative of quantitative limitations (e.g., physical impairment) on mortality (Gurland et al. 2014). Quality-of-life may be conceptualized as multidimensional, with subjective well-being specified as one of the components (Lawton 1997; Seligman and Csikszentmihalyi 2000). The concept of subjective well-being includes the evaluative component of life satisfaction (Diener et al. 1985, 1999; Dolan et al. 2008; Kahneman and Kreuger 2006; Kahneman et al. 2006), experiential or hedonic well-being such as positive and negative affect, and eudemonic constructs such as wellness, self-worth, control, autonomy, self-realization, pleasure and self-activation (Kapteyn et al. 2015).

Conceptual Orientation

Positive affect (PA; Watson et al. 1988) has been defined as a state in which a person is enthusiastic, active and alert, with pleasurable engagement. Low PA connotes sadness. Negative affect (NA) is often measured with depression scales and connotes distress and may include negative mood states. Positive and negative affect are measured at a specific time point with items such as, “Did you feel happy yesterday” or “Are you feeling happy today”. Positive and negative affect items have been found to load on two factors in factor analyses, and have been viewed as different – not just the opposite of one another. However, an alternate view is that the negatively worded items produce a measurement artifact.

Subjective well-being as measured by happiness and satisfaction with life has been studied across the life span in many countries (Blanchflower and Oswald 2008). Other studies (Stone et al. 2010) have examined age differences in negative affect, and positive affect has been found to relate to biomarkers, including reduced inflammatory markers in older women (Steptoe et al. 2012). However, surveys conducted across many countries did not sample disabled elderly persons in the community or those in institutions (Steptoe et al. 2015), the population studied in these analyses.

Aims of the Analyses—The focus of this paper is on methodological issues that arise in the measurement of quality of life domains that include qualitative expressions of positive and negative affect. Specifically emphasized are methods to examine measures that include both positively and negatively worded items as well as methods used to examine the performance of such measures across groups that differ in ethnic and racial composition.

An illustrative example of applications of these methods is provided through the analyses of The Feeling Tone Questionnaire (FTQ; Toner et al., 1999). The FTQ is a measure that was designed to assess affect with items that have been used subsequently in measures of positive affect, e.g., happiness, as well as items traditionally appearing in measures of negative affect such as feeling lonely. The intent was to measure present affective state with items with positive and negative content among individuals with multiple chronic conditions, including communication and cognitive deficits, the majority of whom are elderly. The presence of such positively and negatively worded items has been found to result in methods effects that complicate measurement of affect. A goal was to determine if a methods factor existed, and if the negatively worded items were informative. Additionally, a major goal was to examine performance of the measure in ethnically diverse groups. Although the FTQ has been used in many large-scale studies, such as those examining specialized care approaches for individuals with cognitive impairment (Holmes et al. 1994); little psychometric research has been published, particularly examining performance among ethnically diverse groups. While the FTQ was developed in the 1990's, it remains one of the few measures of affect that can be administered easily among people with comorbidity, communication disorder and cognitive impairment.

Methodological Issues and Methods Effects

Methods effects generally refer to item responses that are affected by factors extraneous to the intended measurement goal, which may result in bias. Examples include acquiescence or yea-saying, social desirability, or a tendency to select more or less extreme item response categories (Bolt and Newton 2011); additionally, rating scales based on agree/disagree continuum responses have been found to be less reliable and valid than those based on item-specific response categories (Saris et al. 2010). For example, different respondents may use the response scale differently. It is assumed that the item response scale in latent variable models is the same for all individuals such that a 5 on a scale for one person means the same thing for other persons. However, this may not be the case due to acquiescence bias or sensitivity to the construct measured, e.g., pain or affect. As reviewed by Maydeu-Olivares and Coffman (2006), variability in the intercepts across participants may require inclusion of an additional factor to model a spurious methodological artifact. Despite the intention that the items measure a single latent construct, negatively worded items may load on one factor and positively worded items on another. Thus, the question arises as to whether there are two subdomains: two constructs of positive and negative affect or simply one construct, affect measured by positively and negatively worded items. From a substantive view, if positive affect and well-being is being assessed, should items measuring health or function such as "pain", "trouble with health" and "anything stopping you from doing what you want to do" be included in the scale?

Measurement errors arising from methodological artifacts may result in inaccurate estimates of traits for certain groups or individuals, manifesting as either over or under estimates of the trait or state, in this instance affect. Such inaccuracies affect comparisons in the context of validity testing or inferential tests of group differences in observational studies or clinical trials. The practice of including positively and negatively worded items together in a scale has been challenged because a change in response set can result in measurement error induced by inattention to a switch in wording or orientation, particularly among those who are frail, suffer from chronic illness, cognitive impairment, or are in palliative care. Method effects have been identified in measures with reverse-scored items (Abbott et al. 2006; Wood et al. 2010). An example in the area of self-reported patient assessments is provided in the analyses of the Patient Reported Outcomes Measurement Information System® (PROMIS®) sleep short forms (Jensen et al. 2016). Response inconsistency associated with reverse-scored items was identified as a methods effect. PROMIS fatigue short form items also showed poor model fit (as manifested by a very low factor loading) when including a reverse-scored item (Reeve et al. 2016).

Several approaches have been used to examine methods effects in the context of positively and negatively worded items. Although there are multiple ways one might consider the impact of methods, constraints, e.g., statistical identification limit these options in practice. Common approaches are to model an uncorrelated method factor on which the negatively worded items load or to model correlated errors between positively and or negatively worded items. The question to be answered is whether or not a methods effect accounting for the negative item wording improves the model fit to the data. Methods effects are examined through confirmatory factor analyses and item response theory (IRT) methods, which produce similar results, depending on parameterization (Reise et al. 1993; Teresi and Jones 2016). However, different traditions are followed in different substantive areas. In the area of patient reported outcomes, such as PROMIS, the approach has been to examine methods effects in the context of model fit to the data and IRT assumption violation. The analyses presented in this paper adopted that approach. However, because there is a large literature on methods related to the study of the construct, affect, more emphasis was placed on the confirmatory analyses examining the dimensionality of the measure prior to examining measurement equivalence.

Measurement Equivalence

Measures of positive and negative affect may perform differently in the context of late life, illness and cognitive impairment. Moreover, ethnicity, language and culture may affect item response. For example, there is an extensive body of research findings related to measurement equivalence in measures of depression, with strong evidence of differential item functioning (DIF; Chan et al. 2004; Cole et al. 2000; Grayson et al. 2000; Kim et al. 2009; Pickard et al. 2006; Yang and Jones 2007). The scale-level impact of DIF has been found to be substantial in some studies (Azocar et al. 2001; Chan et al. 2004; Cole et al. 2000; Kim et al. 2002). (For a review see Teresi et al. 2008.) Carefully constructed depression item banks and short forms (Choi et al. 2010; Pilkonis et al. 2011), such as those developed for PROMIS (Cella et al. 2007; Reeve et al. 2007) have been found to have minimal impact from DIF (Teresi et al. 2009; Teresi et al. 2016). A focus of these analyses is examination of DIF related to ethnic and racial groups in the items from a quality-of-life measure of affect.

METHODS

Sample

The analytic sample, combined from multiple studies included 6,756 individuals, of whom 4,960 were non-Hispanic White, 1,144 non-Hispanic Black and 517 Hispanic (see Table 1). The remainder: Asian (1%) and other (0.6%) was excluded from the analyses. Seventy-three percent were female. The average age was 82 years (s.d. = 11.0); the range in age was from 20 to 107; only 7% were under the age of 65, 2.5% were under the age of 55 and 1.5% under age 50. The average age ranged from 77 among non-Hispanic Blacks to 84 among non-Hispanic Whites. Self-reported education was obtained from several sources including direct assessment and chart review. The direct assessment items were: “How far did you go in school; How much schooling did you have? How many years of formal education did you receive.” The average educational level was 9.4 (s.d. = 5.2) years, including 14% with no formal education. The average educational level ranged from 6.6 years among Hispanics to 9.9 years among non-Hispanic Whites (see Table 1). Race/ethnicity was measured through a combination of chart review and direct assessment. For direct assessment, participants were asked their ethnic/racial group and whether or not they were of Hispanic origin. The questions were: How would you describe your race (White; Black or African American; Asian; Native Hawaiian, or other Pacific Islander; American Indian or Alaska Native; Other)? Are you of Hispanic or Latino descent? What ethnicity do you consider yourself?

The sample was from 17 studies in four types of long-term support services settings: community-residents in home care or adult day care; long-term care residents in assisted living or nursing homes; 757 were in the community; 402 in assisted living and 5,597 in nursing homes. All individuals were receiving some level of care, including palliative care due to the presence of chronic conditions resulting in disability. Cognition was assessed using a variant of the standardized Mini-Mental Status Examination (Molloy and Standish, 1997). About two thirds were cognitively impaired. About one fourth (26%) of the total sample was mildly impaired, one fourth (24%) moderately impaired, and 16% severely impaired. All studies used the same measures and methods for identifying cognitive impairment.

Measure

The Feeling Tone Questionnaire (Toner et al. 1999) is a measure of affective quality-of-life comprised of 16 items, 9 positively and 7 negatively worded items. For the purpose of these analyses, the negatively worded affect items were recoded in the positive direction, such that the resulting scales reflected positive affect. The original measure also included 16 global ratings associated with each item. For these analyses, only the direct response items were examined; these were answered by the respondents, rather than by raters. The three ordinal response categories were no, equivocal (it depends, yes and no) and yes. The measure was designed to include simple statements and response options suitable for administration to individuals with cognitive impairment. Originally, it was posited that quality-of-life in institutional and home care settings could be defined by small events and the quality of interactions with others in a confined, captive and often contentious environment. In such settings, treatment by others takes on greater saliency in defining quality of life, including subjective well-being.

The development samples were of patients selected at random from a sample of 6 facilities from among a probability sample of 25 long-term care institutions in New York and London. Reliability and validity of the measure was examined in the development sample, and in a sample of psychiatric inpatients. Psychometric properties for the original measure were described in Toner et al. (1999). The Cronbach's (Cronbach 1951) alpha was estimated as 0.91 for the long-term care sample and 0.90 for the psychiatric hospital sample. Interrater reliability for two raters of ten cases was estimated as 0.99. Test–retest reliability estimates based on ten cases with a 1-day to 2-day interval between trials was 0.81. The measure correlated 0.30 with a mood scale and 0.71 with physicians' Diagnostic Statistical Manual III-R ratings of depressive symptoms (Toner et al. 1999). In another study of a probability sample of nursing home residents, the interrater reliability for the response scale was estimated as 0.76, and the convergent validity estimates with psychiatric ratings ranged from 0.47 to 0.63 (Teresi et al. 2000).

Analyses

Exploratory analyses were conducted using the first of two random halves of the sample of respondents, as well as on different ethnic/racial subgroups. These analyses were conducted for the total item set and for subsets of items. Initial confirmatory dimensionality analyses were conducted on the second random half of the sample and for different subgroups.

Exploratory Tests of the Model Assumption of Unidimensionality—Item response theory assumptions include unidimensionality and local independence. The latter implies that the items are independent, conditional on the trait level. Model assumptions and fit were tested. Unidimensionality was examined with a) exploratory factor analyses with principal components estimation with cross-loading permitted; and b) confirmatory and bifactor analyses. Results are shown in Table 2.

Permuted parallel analysis (Buja and Eyuboglu 1992; Horn 1965) was performed multiple times to generate a permutation distribution for eigenvalues under the assumption of no association. The observed eigenvalues were then compared to the permutation distribution, and a *p*-value for each eigenvalue obtained. One would expect eigenvalues of 1 in a principal components analysis (PCA) of random uncorrelated data. The number of component eigenvalues significantly greater than and less than 1 provides evidence regarding whether the item set is unidimensional. Parallel analyses were performed using both PCA and factor analyses. Parallel analyses using polychoric correlations (performed with R software, fa.parallel with polychorics; Rizopoulus 2009) have been found to be more accurate than when estimated with other correlation coefficients (Garrido et al. 2012; Green et al. 2016).

Confirmatory Models—Several models were tested to examine explicitly the question of whether one unidimensional affect measure or separate positive and negative measures were statistically superior. The models tested below include those tested by investigators examining measures with positively and negatively worded items (Lindwall et al. 2012; Maydeu-Olivares and Coffman 2006) and of quality-of-life (Chen et al. 2006).

Unidimensional models (labeled models 1a – 1d, Table 3) and models with correlated uniqueness terms (labeled models 2a, 2b, Table 3): Essential unidimensionality was examined through a merged exploratory factor analysis (EFA) and confirmatory factor analysis (Asparouhov and Muthén 2009) performed by fitting a unidimensional model with polychoric correlations using Mplus (Muthén and Muthén 2011). Individual data were input and declared categorical, resulting in estimation using polychoric correlations. Essential unidimensionality has been defined (Stout 1990) as indicating that one dominant latent dimension influenced item responses (see also Bonifay et al. 2015; McDonald 2000; Reise and Haviland 2005). In this formulation, minor local dependencies are permitted. The assumption is that the residual covariances are small, but not zero.

The confirmatory analyses of the unidimensional model specified that all 16 items load on one factor. Additionally, several unidimensional models based on subsets of the items were examined: 13 items, excluding three health-related items, and positive and negative item factor models. The four models were examined for the total sample as well as separately within non-Hispanic Black, Hispanic and non-Hispanic White subsamples. Evaluation of the models was performed with the Comparative Fit Index (CFI; Bentler 1990) and root mean square error of approximation (RMSEA). Cutoff values for model fit in the context of dimensionality (Cook et al. 2009) and invariance (Meade et al. 2008) have been examined critically. Various rules of thumb exist, for example, CFI > 0.95 and RMSEA < 0.08 for adequate fit. Following Cook et al., while these values were used as a guide in this set of analyses; sensitivity analyses with multiple models including bifactor models were conducted to inform the final decisions regarding recommendations for use of the item set. Additional evidence in the form of corrected item-total correlations, reliability and IRT-estimated information was also examined.

A second set of unidimensional models allowed for correlated “residuals”, in which correlations are modeled among the residuals, also known as measurement errors and uniqueness. Model 2a specifies correlated uniqueness for the negatively worded items and 2b correlated uniqueness for both positively and negatively worded items. It is argued that residual variance includes a specific factor in addition to error variance, and should not just be described as error (see Meredith 1993; Meredith and Teresi 2006). Specific factors (sometimes called unique factors) are one of two residual sources of variation within a particular affect item after the influence of common factors (e.g., affect) have been eliminated. Measurement error and systematic individual differences are referred to as uniqueness or a unique factor because they are unique to each item and cannot be separated easily; however, specific or group factors in a bifactor model may capture systematic individual differences common to a set of items in a group factor (Maydeu-Olivares and Coffman 2006). These models (labeled 3 and 4, Table 3) are discussed below.

Methods factor models (model 3): A third set of models examines methods factors. Specific latent method effect factors underlying items of the same method (positively or negatively worded item format) are included together with the latent substantive factor. A bifactor model is specified in which one group factor contains the negative items (the positive items are the reference group); the group factor with negative items is the method factor (see Maydeu-Olivares and Coffman 2006). (The general bifactor model is described below.) This approach captures residual covariation due to the method of including negatively worded items in an affect measure. Model 3 specifies specific latent methods factors for the negative items. Using this method, uncorrelated methods factors are estimated and the item responses are specified to load on one general factor and one group factor representing the negatively worded items; an orthogonal solution is specified and all factors are modeled as uncorrelated.

General bifactor model (model 4): In the bifactor model, a general factor is specified and two additional factors, commonly referred to as group factors are used to model the residual covariation among the items that is not captured by the general factor (Reise 2012; Reise et al. 2007). One additional factor accounts for the residual covariation among the positive items, whereas the second group factor accounts for the residual covariation among the negative items. It is assumed that a single general trait explains most of the common variance but that group traits explain additional common variance for item subsets (Reise et al. 2010). Model 4 is a modified bifactor model in which positive items are specified to load on the first group factor and negative on the second using orthogonal rotation and polychoric correlations estimated in Mplus (Muthén and Muthén 2011).

Additionally, other bifactor models were tested based on the results of a Schmid-Leiman (S-L; 1957) transformation using the “psych” R package (Rizopoulus 2009). All items were specified to load on the general factor, and the loadings (lambdas) on the group factors were specified following the S-L solution (Reise et al. 2010). Mplus (Muthén and Muthén 2011) was used to both estimate the polychoric correlations (based on the underlying continuous normal variables) and to perform the final bifactor modeling and parameter estimation. The bifactor models were compared with a unidimensional model.

Correlated Factors (Model 5): Model 5 specified two correlated factors in which positively worded items were constrained to load on one factor and negatively worded items on the second factor.

Dimensionality Index: The explained common variance (ECV) provides information about whether the observed variance covariance matrix is close to unidimensionality (Sijtsma 2009), and is estimated as the percent of observed variance explained (Reise 2012; Reise et al. 2010). The ECV is the percent of variance explained by the first eigenvalue and was estimated as the ratio of the general factor eigenvalue to the sum of the general and group factor eigenvalues.

Item Response Theory Models—IRT models were used to examine item- and scale-level DIF, including magnitude and impact (see below for a definition).

Local Dependence (LD): An assumption of IRT is local independence. This assumption was tested using the generalized, standardized local dependency chi-square statistics (Chen and Thissen 1997) provided in IRTPRO, version 2.1 (Cai et al. 2011). Because LDs are affected by sample sizes, smaller random samples of 300 were also used in sensitivity analyses.

IRT-model Fit: The root mean square error of approximation from IRTPRO (Cai et al. 2011) was used to assess IRT model fit.

Anchor Items and Linking: DIF-free anchor items were used to set the metric and link the comparison groups on affect. The mean and variance for the target groups studied were estimated, and the reference group mean and variance was set to 0 and 1, respectively. Several anchor item methods recommended for use have been reviewed recently (Kopf et al. 2015; Setodji et al. 2011; Teresi and Jones 2016; Wang et al. 2012; Woods 2009). An iterative modified "all-other" anchor method (Orlando-Edelen et al. 2006; Thissen et al. 1993) was used in selection of the anchor items for theta estimation in these analyses. Initial DIF estimates were obtained by treating each item as a "studied" item, while using the remainder as "anchor" items. The analyses were repeated using the final subset of items identified as free of DIF as the "purified" anchor set. Items with DIF from the original anchor set were removed.

Model for DIF Detection: The graded response model (Samejima 1969) was used for the analyses of DIF. An item shows DIF if people from different subgroups but at the same level of the attribute (denoted theta or Θ) have unequal probabilities of endorsement. The item characteristic curve (ICC) that relates the probability of an item response to the underlying state, e.g., affect, measured by the item set is characterized by: a discrimination parameter (denoted a) and location (severity) parameter(s) (denoted b). The presence of DIF is demonstrated by ICCs that are different for the subgroups examined.

DIF Detection Tests: The Wald test for examination of group differences in IRT item parameters was the primary analytic method; the Wald test is an expansion of Lord's chi-square tests for DIF (Lord, 1980; Teresi, Kleinman et al. 2000; Woods et al. 2013). For each studied item, a model was constructed with all parameters (except the studied item) constrained to be equal across comparison groups for the anchor items, and item parameters for the studied item freed to be estimated distinctly. An overall simultaneous joint test of differences in the a or b parameters was performed followed by step down tests for group differences in the a parameters, followed by conditional tests of the b parameters. Uniform DIF (defined as DIF in a constant direction across the trait) was detected when the b parameters differed and non-uniform DIF (defined as DIF in different directions at different points along the latent affect continuum) when the a parameters differed. Given that the interest was in comparing the studied groups to the reference group, non-orthogonal contrasts were used. The final p values were adjusted using the Bonferroni (1936) method for adjustment for multiple comparisons. Other methods recommended by Thissen, Steinberg and Kuang (2002) such as Benjamini-Hochberg (B-H; Benjamini & Hochberg, 1995) typically yield very similar results.

Sensitivity Analyses for DIF Detection: A second DIF-detection method used in sensitivity analyses was based on ordinal logistic regression (OLR; Swaminathan and Rogers 1990; Zumbo 1999). However, a variant of this model conditions on the latent variable derived from IRT: IRTOLR (Crane et al. 2006; Crane et al. 2004; Mukherjee et al. 2013). The affect estimates from a graded response IRT model were used as the conditioning variable, and effect sizes were incorporated into the uniform DIF detection procedure. Uniform DIF is defined in the OLR framework as a significant group effect, conditional on the affect state and non-uniform DIF as a significant interaction of group and affect state. Three hierarchical models were tested; the first examines affect state (1), followed by group (2) and the interaction of group by state (3). Non-uniform DIF is tested by examining model 3 vs. 2; uniform DIF is tested by examining the incremental effect of model 2 vs. 1, with a chi-square (1 degree of freedom) test (Camilli and Shepard 1994). The software, lordif (Choi et al. 2011) was used to perform IRTOLR.

Evaluation of DIF Magnitude and Effect Sizes for Primary Tests of DIF: The magnitude of DIF refers to the degree of difference in item performance between or among groups, conditional on the trait or state, operationalized as differences in expected item scores. An expected item score is the sum of the weighted (by the response category value) probabilities of scoring in each of the possible categories for the item. The method used for quantification of the difference in the average expected item scores was the non-compensatory DIF (NCDIF) index (Raju et al. 1995) used in DFIT (Oshima et al. 2009; Raju 1999; Raju et al. 2009). An additional effect size measure denoted T1, proposed by Wainer (1993) and extended for polytomous data by Kim et al. (2007) was also examined (see also Kleinman and Teresi, 2016). Cutoff values for magnitude were established based on simulations (Fleer, 1993; Flowers et al. 1999). For example, the cutoff value recommended by Raju is 0.024 for polytomous items with three response options (Raju, 1999); this cutoff corresponds to an average absolute difference greater than 0.155 (about 0.16 of a point) on a three point scale.

Evaluation of DIF Impact: Expected scale score functions (also referred to as the test or scale response function) are the sum of expected item scores, and were examined as evidence of aggregate-level impact. The effect of DIF on the total score was examined by calculating group differences in the test response functions; these differences provide overall aggregated measures of DIF impact.

Information: Finally, the item and test information functions from IRT were calculated and graphed. These curves are useful in evaluation of items because non-informative items are indicative of items that do not discriminate well and are not related well to the affect state measured. In the current context, the information functions also served to identify a shorter-form of the measure.

Evaluation of Reliability—Although Cronbach’s alpha (Cronbach 1951) was calculated, ordinal alpha based on polychoric correlations (Zumbo et al. 2007) was also estimated. Examining Cronbach’s alpha alone does not provide adequate evidence regarding the performance of the scale because it lacks the invariance properties of statistics derived based on parameters estimated from factor, regression and IRT analyses. As such, Cronbach’s alpha cannot be compared legitimately among groups. Other methods based on latent variable models are preferable (Zinbarg et al. 2005). Ordinal alpha estimates appropriate for ordinal data are based on polychoric correlations and correspond better with estimates from latent variable models. Polychoric correlations assume an underlying latent response variable, and are thus more invariant with respect to marginal distributions of response categories (base rate). The correlation between two variables (items) can be represented as a path model with the true score (latent trait) causing the two item response variables (Uebersax 2000; Lord and Novick 1968).

Other reliability estimates that are based on latent variable modeling include McDonald's (McDonald 1999) omega total (ω_t), estimated from a factor model. This reliability estimate is based on the proportion of total common variance explained. Because McDonald's omega is typically derived from a latent bifactor model, it is arguably more invariant than values based on observed response models (see also Bentler, 2009). Software to calculate these indices, developed by Revelle and Zinbarg (2009) are contained in the "Psych" package in R (Revelle 2015; www.R-project.org; R Development Core Team 2008). Additionally, IRT-based reliability statistics were examined at selected points along the underlying latent continuum (theta). These conditional reliability estimates were based on the definition of reliability as 1- the ratio of error variance to total variance, and operationalized by subtracting from 1 the weighted squared standard error of theta at selected theta values (see for example, Cheng et al. 2015; Teresi et al. 2000). These relationships can also be presented in the context of IRT information (Cheng et al. 2015).

RESULTS

Exploratory Tests of Dimensionality

Exploratory Principal Components Analyses—The estimate of the ratio of the first component to the second for the 16 item set for the first random half and the total sample was 2.4 and 2.3, and the first component explained 32% to 33% of the variance (see Table 2). The scree plot for the total sample is given in Appendix, Figure 1. Based on the parallel analyses with polychoric correlations (not shown), it appeared as if more than one component should be retained. The parallel analyses results, whether estimated using PCA or factor analyses suggested the retention of more than one factor (eigenvalues for the original factors were greater than those of the simulated data). This result provides evidence that essential unidimensionality for this item set may be in question. The 16 item set includes the health-related items: "Do you have any pain?"; "Is there anything that stops you from doing what you want to do;" and "Do you have any trouble with your health". Although somatic items have been included in measures of affective state, they are more generally related to depression, e.g., appetite and sleep. Preliminary analyses suggested that the three items related to health loaded on a separate factor. Thus, analyses were also conducted on a 13 item set removing the three items with health-related content.

The analyses conducted among the different racial/ethnic groups showed that the 16 item set treated as unidimensional did not perform as well among non-Hispanic Black and Hispanic groups (see Table 2). Ratios varied across ethnic/racial groups for the 16 item set, and were lower for non-Hispanic Blacks and Hispanics. Ratios were more similar and close to 3 for all race/ethnic groups for the 13 item set.

Finally, analyses were performed separately with the 9 positively and 7 negatively worded items. The ratio was higher (5.3) for the 9 positive affect items and the first component explained 51% of the variance. The ratio for the 7 negative affect item set was low at 2.4 and the first component explained 39% of the variance. Ratios were all above 4 for the positive item set (4.3 for Hispanics to 5.4 for non-Hispanic Whites). The negative item set ratios were lower, ranging from 2.2 for non-Hispanic Blacks to 2.7 for Hispanics.

Confirmatory Factor Models Examining Methods Effects

Model 1; Unidimensional Model: Table 3 shows the results for four models: 16 items (model 1a), 13 items (model 1b, with three health items removed), and positively and negatively worded items (models 1c, 1d), all specified as unidimensional. The results of the unidimensional CFA show lower item loadings on the underlying factor for all negative affect items compared to the positive affect items. The loadings for the following health-related items were under 0.30: “Do you have any pain?”; “Is there anything that stops you from doing what you want to do?”; “Do you have any trouble with your health?” The loadings for the 13 item set (model 1b) changed only between 0.01 to 0.04 points; the four negatively worded items all evidenced lower loadings (0.29 to 0.35). The analyses for the 16 and 13 item sets using the second random half of the sample were consistent with those reported above.

When analyzed separately, the item loadings within the positive and negative item sets were consistent and within an acceptable range. The loadings for the positive affect set (model 1c) ranged from 0.60 to 0.75; while the loadings for the negative affect set (model 1d) ranged from 0.44 to 0.61. The highest loading among the negatively worded items was 0.61 (Item 15, “Do you have any trouble with your health”).

Appendix Table 1 is a summary of the fit statistics for the models described above, and several others tested. The model fit indices (CFIs) for the unidimensional CFA from Mplus for the 16 item set was 0.829 for the total sample and ranged from 0.677 for Hispanics to 0.848 for non-Hispanic Whites. The CFI for the 13 item set increased for the total sample and the subgroups from 0.829 for the 16 item version to 0.925 for the total sample; for non-Hispanic Whites, the CFI was 0.928 and 0.923 for non-Hispanic Blacks. However, the CFI for Hispanics was lower (0.882), perhaps due to the unequal sample sizes and lower sample size for Hispanics. The improvement was seen also in the RMSEA statistics which changed from 0.105 for the 16 item set to 0.084 for the total sample for the 13 item set, and ranged from 0.081 for White to 0.106 for the Hispanic subgroups (see Appendix Table 1).

Model 2; Correlated Uniqueness: As shown in Table 3 (16 item set) and Appendix Table 1, the model fit statistics improved for models 2a and 2b (correlated negative uniqueness and negative and positive uniqueness); the health-related items loaded the lowest (<0.30) as in model 1. The CFI was 0.940 for model 2a with correlated negative uniqueness for the 16 item set, and 0.963 for the 13 item set (Appendix, Table 1); the RMSEA estimates were 0.070 and 0.062, respectively. The CFI values for model 2b were 0.973 and 0.994 for the 16 and 13 item set, respectively; the RMSEA values were 0.061 and 0.039; indicating better fit for the 13 item measure.

Model 3; Methods Factors: Model 3 included a latent methods factor, a bifactor model where all items were specified to load on the general factor and the negative affect items on a separate factor. All the negative affect item loadings on the latter (methods) factor were higher compared to their loadings on the general factor, and ranged from 0.34 to 0.62, while loadings of the same items on the general factor ranged from 0.15 to 0.32. In contrast, the positive affect item loadings on the general factor ranged from 0.61 to 0.75 (See Table 3, model 3). The three health related items loaded highest on the methods factor.

Model 4; Bifactor Model: The two group bifactor model shows a clear dominance of a positive group factor and a less well-defined negative factor, dominated by the health items. Very low loadings were observed for non-health negatively worded items on the negative group factor. The general factor loadings varied from that of the other solutions, indicating model misspecification, when all items were included in the model.

In contrast, a bifactor model of the 13 item set, which was used as an additional test of dimensionality showed that the loadings on the single common factor were very similar to those observed on the general factor from the bifactor analyses (the range of differences was from 0.01 to 0.08), which provides some evidence for acceptable unidimensionality after removal of the health related items (loadings not shown). The ECV dimensionality statistic increased from 26.337 to 68.706 for the total sample, and was 34.049 for non-Hispanic Black, 59.743 for Hispanic and 70.614 for White subgroups for the 13 item set (see Table 4). The above results of the 13 item set analyses lend support for essential unidimensionality as a condition for the differential item functioning analysis; however, the ECV values were lower for non-Hispanic Blacks.

Model 5; Two Factor Model: Specifying a two factor model for the positively and negatively worded items with oblique rotation fit equally as well as the methods factor model (model 3); however, the CFI for both models were lower than desirable (0.926, 0.920). Model 5 yielded reasonable loadings for both factors (0.61 to 0.75 for the positive items, and 0.51 to 0.62 for the negatively worded items). These results were similar to models 1c and 1d, treating the positively and negatively worded items as unidimensional (see Table 3).

A three group bifactor model was also tested in which the health items were specified to load on a third group factor; however, the loadings on the general and first group factors were inconsistent with all other models.

Summary—As is evident from the above results, a 16 item set modeled as unidimensional did not fit well, and was the worst fitting model across subgroups. The models with correlated uniqueness (models 2a and 2b, Table 3) fit best among the unidimensional models for both the 16 and 13 item sets, indicating that the negative item wording may be inducing a methods artifact. The 13 item solutions with correlated uniqueness (Appendix, Table 1) were superior to both the 16 item correlated uniqueness solutions, and the 16 item bifactor model (model 4). Although the two factor solution with oblique rotation appeared to support a positive and negative factor, the negative affect item set did not fit well across race/ethnic groups with CFIs ranging from 0.875 to 0.892 (see Appendix Table 1). The 13 item unidimensional model was a better fit, and the positive item solution fit best among the unidimensional models (CFI=0.967). The positive affect item set solution fit relatively well across most ethnic/racial subgroups (CFI=0.939 to 0.968). Thus, although one may argue about how to proceed, given the goal of determining if the measure was essentially unidimensional, and the original intent of the scale to measure one underlying dimension; DIF testing was performed with the 13 and 9 item set.

Reliability Estimates

The corrected item-total correlations (not shown) for the 16 item scale ranged from 0.26 to 0.52, from 0.24 to 0.55 for the 13 item measure, from 0.45 to 0.58 for the positively worded, and from 0.24 to 0.38 for the negatively worded items. The Cronbach’s alpha estimates were 0.77 and 0.78, respectively for the 16 and 13 item measure (see Table 4); however, the ordinal alpha estimates using polychorics were 0.85 and 0.86, respectively, similar to McDonald’s omega total. The McDonald’s omega estimate was 0.86 for the total sample for the 13 item set. The reliability estimates for the demographic subgroups were in the same range. The estimate for ordinal alpha for both the Hispanic and non-Hispanic Black subsamples was 0.83 for the 16 item scale, and the McDonald’s omega was 0.84 for both. The ordinal alpha for the 13 item scale for the Hispanic subsample was 0.84, and 0.85 for the non-Hispanic Black subsample; the McDonald’s omega was 0.85 for both groups.

The omega total values for the positive item set (Table 4) ranged from 0.87 to 0.88, and from 0.72 to 0.79 for the negative affect item set. The negative item set evidenced lower values (in the 0.70’s) across subgroups defined by race and ethnicity.

Finally, the reliability estimates (precision) at points along the latent positive affect trait (theta) reflective of where respondents were observed ranged from 0.69 to 0.90 for the 13 item set (Table 5) across ethnic/racial groups. The highest reliability values were in the theta range of −2.0 to 0.4 and the lowest were at thetas from 1.2 to 1.6, reflective of higher positive affect. The overall reliability estimate was 0.85 or 0.86 across groups. Examining the positive and negative item sets (Table 6), the positive item set evidenced relatively high average reliabilities across groups, and across levels of theta except at the highest positive affect level; the values for the negative item set were lower across subgroups, ranging from 0.73 in the total sample and non-Hispanic White and non-Hispanic Black samples to 0.78 among Hispanics (see Table 6).

IRT Parameter Estimates, Tests of DIF and Assessment of Magnitude and Impact

Local Independence Assumption: Most of the LD chi square statistics (not shown) for the 13 item set were within an acceptable range, however values for a few item pairs were higher: Item 10 – “Do you like people here?” with Item 6 – “Are people helpful here?” (LD = 28.4); Item 12 – “Do you sleep well?” with Item 9 – “Do you have a good appetite” (LD = 27.2) and Item 1 = “Are you feeling well?” (LD = 24.5).

Item Parameter Estimates: Shown in Table 7 are the graded response item parameters and their standard errors for the total sample. The category response functions (not shown) were non-overlapping and showed an ordinal pattern for the middle category. Appendix Tables 2 – 4 show the discrimination (*a)* parameters across subgroup comparisons. As shown, the *a* parameters varied somewhat across items and groups, ranging from 0.50 (lonely) to 2.30 (nice day yesterday) for the 13 item set for the total sample. The range across ethnic/racial subgroups was similar. Additionally, discrimination parameters were similar across versions; although parameters were higher for most negatively worded items when examined as a separate item set, they remained lower than the positive items (see Appendix Table 2). As expected *b* parameters were remarkably stable across item sets (see Appendix Tables 3 and 4).

IRT Information, Scale Means and Association Among the Scales: Also examined were the information functions (from IRT) for the items and scale scores. At the item level, regardless of what item set was examined: 16, 13, negative items alone, the negative items did not provide much information, relative to the positively worded items. As shown in Figures 1 and 2, the most informative items were: “nice day yesterday”, “like people here”, “people are helpful here” and “happy”. The least informative items were the negatively worded items. This result is also shown in Figure 4 where it was observed that the negatively worded items provided low information when treated as a separate set; whereas Figure 3 shows the relatively higher information provided by the positive items.

Examining the test (scale) information function, for the 13 and positive item set, the peak is at theta = −0.4. For the negative item set the information peak is at theta = −1.2. The 13 and 9 item scales provided overall peak information between 8 and 9; whereas the negatively worded items provided much less information overall with peak information of about 3. Similar to the 13 item set, the positive affect item set composed only of positive items performed relatively well in terms of information, providing most information between theta levels of −2.0 to 0, in the lower range of the affect scale.

The correlation of the 13 and 9 item scales (using theta score estimates) was 0.94; however, the estimated correlation of the negative affect scale with the 13 item scale was 0.55, and 0.32 with the positive item scale.

DIF Results—DIF analyses were performed for racial ethnic subgroups for the 13, and positive affect item sets. Summary DIF results are presented in Tables 8 and 9, and details are presented in Appendix Tables 3 and 4. As expected with larger samples, significant DIF after Bonferroni adjustments was observed for many items.

After removing the three health related items, the DIF analysis was performed for the 13 item set. The items “nasty” and “feeling good about tomorrow” were identified as anchor items. The third item, “like being here”, showed DIF in the anchor identification procedure; however, it did not evidence DIF in the final DIF analyses. Consistent DIF across methods was observed for six items: “happy”, “helpful”, “lonely”, “appetite”, “like people”, “upset yesterday”.

One anchor item was identified initially in the 9 positive affect item analysis: “Do you like being here?” Consistent DIF across methods was observed for three items: “happy”, “appetite” and “feel good about tomorrow”.

Conditional on positive affect, non-Hispanic Black as contrasted with non-Hispanic White respondents evidenced a lower probability of endorsing the items, “feeling lonely” and “bored”. (It is noted that the curve for non-Hispanic Blacks is above the curves for the other groups in the graph in Figure 5 because of reverse scoring.) Hispanic as compared to non-Hispanic White respondents evidenced a higher probability of endorsing the item, “happy” across all analyses. The item, “sleep well” was significantly more discriminating for the Hispanic respondents in the positive affect item set analyses. No item had significant non-uniform DIF after Bonferroni correction in the 13 item set analysis.

Magnitude: Although, many items were flagged with DIF, the magnitude was small. The item, “feeling lonely” evidenced an NCDIF statistic above the threshold for the comparison of non-Hispanic Black with non-Hispanic White respondents for the 13 item set analyses. Within the positively worded item set, no items evidenced DIF of high magnitude by the primary NCDIF criterion.

Sensitivity Analyses: Sensitivity analyses were conducted to examine the effect of increasing the size of the anchor item set on the results. Because of large sample sizes, significant DIF was observed for many items. Final DIF analysis for the 13 and 9 item sets included two anchors. For both sets, the DIF analysis was repeated by including four anchor items each. The log likelihood statistic rank order method was employed for the selection of the items. There were two changes in the 13 item set analysis: the items, “did anything upset you yesterday” and “sleep well” changed to significant after the Bonferroni correction for the Hispanic group in comparison to non-Hispanic Whites. There were no changes for the 9 positive affect item set with the inclusion of the anchor items, “like being here”, “people are helpful”, “like people here”, and “feel good about tomorrow”. Although there was evidence of sensitivity of the DIF findings to the number of anchor items, the results are equivocal because increasing the number of anchor items introduces some DIF into the anchor set, which can result in false DIF detection.

Aggregate Impact: As shown in Figure 5, there was no evident scale level impact. All group curves were overlapping for all comparisons. Examining the average model-based trait mean (theta) estimates for the 13, and positive item sets, it was observed that after DIF adjustment (estimating parameters separately for each group for items evidencing DIF) non-Hispanic Blacks evidenced slightly lower affect scores (mean = −0.13, −0.19 for the 13 and nine item sets) in comparison to the White reference group (with the theta estimate set to 0 for the reference group). The Hispanic averages were even lower (−0.30). Prior to DIF adjustment, differences were somewhat larger between non-Hispanic Whites and Blacks (−0.23 and −0.26) and between non-Hispanic Whites and Hispanics (−0.32 and −0.35) for the 13 and nine item sets.

DISCUSSION

The focus of these analyses was to illustrate methodological approaches for examining potential methods effects associated with positively and negatively worded items. A goal of the analyses was also evaluation of the performance of a measure of affect across racial and ethnic groups. A methodological question posed was whether the measure was unidimensional with positively and negatively worded items or if two constructs, positive and negative affect were being assessed. It was posited that a methods factor might be present, and various models were tested to determine if there were two viable factors or if one construct was underlying the item intercorrelations, with a method effect inducing loadings of negatively worded items on the second factor.

As is evident from the results, a 16 item set modeled as unidimensional did not fit well; the health-related items loaded below 0.30. Removal of the three physical health-related items appearing to measure a different construct resulted in improved fit. The 13 item model with correlated uniqueness fits best among the unidimensional models, indicating that the negative item wording may be inducing a methods artifact. All the negatively worded affect item loadings on the methods factor were higher compared with their loadings on the general factor, and ranged from 0.34 to 0.62, while loadings of the same items on the general factor ranged from 0.15 to 0.32 (results not shown). In contrast, the positive affect item loadings on the general factor ranged from 0.61 to 0.75. What is clear from model 3 is that the three health-related items loaded highest on the methods factor, again providing evidence that these items appear to be measuring a different construct.

The results of the bifactor model analyses revealed the presence of two possible group (secondary) factors: one with positively worded and one with negatively worded items. However, the "pain" and "health" items appeared to load strongly on a secondary factor, providing further evidence that these health-related items do not fit with the others. They are also not a good fit, conceptually. An alternative approach would be to treat the negatively worded items as a separate measure. However, the results of the analyses of the negatively worded affect measure demonstrated that little variance in the item set was explained (eigenvalues of 2.7 and 1.1) and poor fit (CFI<0.9) and lower loadings were observed for the negatively as contrasted with positively worded items (range of 0.44 to 0.61 versus 0.60 to 0.75). These items were also less informative, and less overall IRT information was provided at the scale level; finally, the negative item set was less reliable.

A comparison of all the models tested showed that the best fitting model was the 13 item model with correlated uniqueness, lending support to the contention that rather than positive and negative affect, positive affect was the predominant construct measured. The inclusion of some negatively worded items in the 13 item version although not as well-performing did not degrade appreciably the overall scale performance.

Although the two factor solution with oblique rotation appeared to support a positive and negative factor, the negative affect item set was not a good fit. The positive affect item set modeled as unidimensional fit well. It is possible that the negatively worded items are measuring a secondary trait or methods factor related to a tendency toward "negativism";

however, the dominance of the health-related items among the negatively worded item set in some of the models may suggest a health-related secondary trait. Given that a goal was to determine if the measure was essentially unidimensional, and the original intent of the scale was to measure one underlying dimension, positive affect; it is argued that a 13 item measure or the 9 item positive affect items would be acceptable scales.

Reverse-scored items are often included to minimize acquiescence-bias. However, evidence for their effectiveness is relatively sparse (van Sonderen et al. 2013). The evidence presented in the analyses of PROMIS sleep (Jensen et al. 2016) and fatigue (Reeve et al. 2016) short forms as well as that presented here adds to arguments that such items may confuse some participants and complicate the interpretation of scores. These findings suggest that investigators should consider whether or not the inclusion of reverse-scored items is necessary because of clinical content relevance, and worth the risk of error due to changes in response direction. These considerations are important in the context of ensuring minimization of measurement error, particularly when such items are administered to frail individuals with cognitive impairment.

Methods effects are often dealt with through modeling, which may be a solution when measures are used analytically; however, such modeling will not be useful in the context of clinical use of a measure in a field setting or in administration of small subsets of items in computerized adaptive testing. An important issue is what approach to take in practice.

Although many items were flagged with DIF, the magnitude was small. Only two items evidenced DIF above the magnitude threshold. The item, “feeling lonely” evidenced an NCDIF statistic above the threshold for the comparison of non-Hispanic Black with White respondents for the 13 item set analyses. Conditional on positive affect, non-Hispanic Black as contrasted with non-Hispanic White respondents evidenced a lower probability of endorsing the item, “lonely”. DIF has been observed for items related to loneliness in other studies, specifically with respect to age (Choi et al. 2009; Estabrook et al. 2015).

Hispanic respondents evidenced a higher probability of endorsement of the item, “happy” compared to the non-Hispanic White respondents across analyses; however, the DIF was not of high magnitude. Differential item functioning has been observed for happiness items with respect to age (Choi et al. 2009; Estabrook et al. 2015); education in the direction of lower education associated with higher happiness, conditional on depression (Perkins et al. 2006); and gender (Yin et al. 2015). In the latter study, conditional on the depression trait, men were less likely to endorse being happy. In comparing positively and negatively worded items, Iwata and colleagues (2002) found that immigrant Hispanics evidenced higher scores on the Center for Epidemiologic Studies Depression Scale (CES-D; Radloff, 1977) than other ethnic groups and non-Hispanic Whites, indicating less positive affect. However, no differences were observed for negatively worded items. The authors conclude that the under-endorsement of positive items might result in greater differences, possibly measurement artifacts among Hispanics born outside the United States. In this study similar results were observed. Hispanic respondents evidenced lower mean affect scores after DIF adjustment (−0.30) than the non-Hispanic White reference group across all item sets; however, the difference was somewhat less for the negatively worded items (−0.23). Studies of other constructs such as self-esteem have documented systematic differences across countries differing in cultural affiliations in responses to negatively worded as contrasted with positively worded items (Lindwall et al. 2012). Systematic methods effects related to personality constructs such as a tendency toward negativism may be operating in scales with such items.

From a methodological perspective, it is argued that both confirmatory factor analytic methods and item response theory modeling are complementary, providing information about item discrimination and information as well as reliability and differential item functioning.

From a clinical perspective, the results conform to the quality of life assessment framework that posits that qualitative aspects of positive affect are different from physical health-related aspects. In the context of restricted environments as experienced by those with disability and comorbidity, applied interventions for preserving, improving and/or achieving quality of life goals can be designed by targeting and reducing restrictions of choice (Gurland et al. 2010; Gurland et al. 2009). The nine item FTQ may provide a shorter measure of affective quality-of-life for use in clinical research among older, ethnically diverse individuals with disability and cognitive impairment.

Limitations

Limitations of the analyses include the inability to examine Hispanic subgroups. Additionally, the equivalence of 6.6 years of formal education among the Hispanic respondents to those schooled in the United States cannot be established. Moreover, educational quality varies; thus the years of education of non-Hispanic Blacks born in the South may not equate to that of Whites. This is a limitation; however, the data are presented descriptively to provide some evidence of the low educational level of this sample. Finally, the paucity of anchor items could have affected the DIF results; however, inclusion of items with DIF in the anchor set can result in false DIF detection. Given the low magnitude of DIF observed and the few changes observed in sensitivity analyses with inclusion of additional anchor items, it is not likely that false DIF detection posed a threat to the conclusions.

Conclusions

These analyses illustrated the use of both a structural equation and item response theory approach to examining the performance of an affect measure. The confirmatory and bifactor models were used to converge upon the identification of methods effects associated with administration of positively and negatively worded items. The IRT methods contributed additional information about magnitude and impact of DIF and information provided by the item sets. Thus the use of both approaches is recommended. Although the models are parameterized in a similar manner, the byproducts are somewhat different. The CFA analyses were focused on parameter estimation for the entire sample, while the IRT analyses focused on parameter comparisons across groups. The first part of the analyses was to examine dimensionality and methods factors. The second part using IRT was to examine DIF. Measurement equivalence could have been examined using CFA; however, DIF magnitude methods have been developed in the IRT environment, and for ease of application, this method was selected. Moreover, in the context of binary and ordinal data, IRT models may be preferable for the invariance stage of analyses (see for example, Sass, Schmitt and Marsh, 2014).

The FTQ analyses were illustrative of how negatively worded items might influence response among older, frail individuals. This will be the case whether or not one is measuring one construct with positively and negatively worded items or if one is measuring positive and negative affect because the latter will usually also have negatively worded items. In other words, many "negative" affect measures will be negatively worded. Thus, one may have similar problems from a methods perspective regardless of the intent. As reviewed in McHorney and Fleishman (2006), older persons have been found to give "rosy" responses, indicative of positive response bias. Some have argued that methods effects associated with negatively worded items may not imply simply a methods artifact but could reflect a latent construct related to "negativity" or to some other personality trait. However, what is shown here is that these negative items were not informative. Some of the negative items, e.g., loneliness have appeared in many well-validated affect measures, and it is thus not likely that the content itself is the issue with these items. Moreover, some negatively worded items, e.g., "anything upset you yesterday" that were less informative were similar in content to the positively worded items, e.g., "have a nice day yesterday" that performed better. Thus, it would appear that it is more likely that a methods artifact is at play. For a review of these issues, see Vecchione et al., 2014.

The substantive findings suggest that a 13 item measure comprised of positively and some negatively worded items fit the data reasonably, and a 9 item positive affect item set fit well. The nine item scale comprised of only positively worded items has the advantage of being shorter, and because it performed somewhat better than the 13 item measure may be preferred. These item sets performed well in terms of IRT information and reliability across groups studied and evidenced little DIF of high magnitude and overall low scale-level DIF impact across non-Hispanic Black and Hispanic groups in comparison to non-Hispanic White older persons. Thus, these scales can be recommended to assess the affective component of quality-of-life among ethnically diverse groups of frail and/or cognitively impaired individuals residing in the community or in institutional settings.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Partial funding for the analyses was provided by the National Institute on Aging (NIA)-funded Mt. Sinai Pepper Center, P30AG028741 (PI: Siu), and the NIA Edward R. Roybal Center, P30AG022845 (PI: Reid, Pillemer, Wethington). The authors thank Stephanie Silver, MPH for editorial assistance in the preparation of this manuscript.

References

Abbott RA, Ploubidis GB, Huppert FA, Kuh D, Wadsworth ME, Croudace TJ. Psychometric evaluation and predictive validity of Ryff's psychological well-being items in a UK birth cohort sample of women. Health and Quality of Life Outcomes. 2006; 4:76.doi: 10.1186/1477-7525-4-76 [PubMed: 17020614]

Albert SM, Teresi JA. The MacMillan Encyclopedia of Aging. New York: MacMillan References, U.S.A; 2002. Quality of life, definition and measurement.

Asparouhov T, Muthén B. Exploratory structural equation modeling. Structural Equation Modeling. 2009; 16:397–438. DOI: 10.1080/10705510903008204

Azocar F, Areán P, Miranda J, Muñoz RF. Differential item functioning in a Spanish translation of the Beck Depression Inventory. Journal of Clinical Psychology. 2001; 57(3):355–365. DOI: 10.1002/jclp.1017 [PubMed: 11241365]

Benjamini Y, Hochberg Y. Controlling for the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, Series B. 1995; 57:289–300.

Bentler PM. Alpha, dimension-free, and model-based internal consistency reliability. Psychometrika. 2009; 74:137–143. DOI: 10.1007/s11336-008-9100-1 [PubMed: 20161430]

Bentler PM. Comparative fit indexes in structural models. Psychological Bulletin. 1990; 107(2):238–246. DOI: 10.1037/0033-2909.107.2.238 [PubMed: 2320703]

Blanchflower DG, Oswald AJ. Is well-being U-shaped over the life cycle? Social Science Medicine. 2008; 66:1733–1749. DOI: 10.1016/j.socscimed.2008.01.030 [PubMed: 18316146]

Bolt DM, Newton JR. Multiscale measurement of extreme response style. Educational and Psychological Measurement. 2011; 71(5):814–833. DOI: 10.1177/0013164410388411

Bonferroni CE. Teoria statistica delle classi e calcolo delle probabilità. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze. 1936; 8:3–62.

Bonifay WE, Reise SP, Scheines R, Meijer BR. When are multidimensional data unidimensional enough for structural equation modeling? An evaluation of the DETECT multidimensionality index. Structural Equation Modeling. 2015; 22:504–516. DOI: 10.1080/107005511.2014.938596

Brod M, Stewart AL, Sands L, Walton P. Conceptualization and measurement of quality of life in dementia: The dementia quality of life instrument (DQoL). The Gerontologist. 1999; 39:25–35. DOI: 10.1093/geront/39.1.25 [PubMed: 10028768]

Buja A, Eyuboglu N. Remarks on parallel analysis. Multivariate Behavioral Research. 1992; 27(4): 509–540. DOI: 10.1207/s15327906mbr2704_2 [PubMed: 26811132]

Cai L, Thissen D, du Toit SHC. IRTPRO: Flexible, multidimensional, multiple categorical IRT Modeling [Computer software]. Chicago, IL: Scientific Software International, Inc; 2011.

Camilli G, Shepard LA. Methods for identifying biased test items. Thousand Oaks, CA: Sage Publications; 1994.

Cella D, Yount S, Rothrock N, Gershon R, Cook K, Reeve B, et al. on behalf of the PROMIS Cooperative Group. The patient-reported outcomes measurement information system (PROMIS): progress of an NIH roadmap cooperative group during its first two years. Medical Care. 2007; 45(5 Suppl 1):S3–S11. DOI: 10.1097/01.mlr.0000258615.42478.55

Chan KS, Orlando M, Ghosh-Dastidar B, Sherbourne CD. The interview mode effect on the Center of Epidemiological Studies Depression (CES-D) scale: An item response theory analysis. Medical Care. 2004; 42(3):281–289. DOI: 10.1097/01.mlr.0000115632.78486.1f [PubMed: 15076828]

Chen WH, Thissen D. Local dependence indices for item pairs using item response theory. Journal of Educational and Behavioral Statistics. 1997; 22:265–289. DOI: 10.2307/1165285

Chen FF, West SW, Soussa KH. A comparison of bifactor and second-order models of quality of life. Multivariate Behavioral Research. 2006; 41:189–224. [PubMed: 26782910]

Cheng Y, Liu C, Behrens J. Standard error of reliability estimates and the classification accuracy and consistency of binary decisions. Psychometrika. 2015; 80:645–664. DOI: 10.1007/s11336-014-9407-z [PubMed: 25228494]

Choi H, Fogg L, Lee EE, Choi Wu M. Evaluating differential item functioning of the CES-D Scale according to caregiver status and cultural context in Korean women. Journal of the American Psychiatric Nurses Association. 2009; 15(4):240–248. DOI: 10.1177/1078390309343713 [PubMed: 21665810]

Choi SW, Gibbons LE, Crane PK. lordif.: An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. Journal of Statistical Software. 2011; 39:1–30. DOI: 10.18637/jss.v039.i08

Choi SW, Reise SP, Pilkonis PA, Hays RD, Cella D. Efficiency of static and computer adaptive short forms compared to full-length measures of depressive symptoms. Quality of Life Research. 2010; 19:125–136. DOI: 10.1007/s11136-009-9560-5 [PubMed: 19941077]

Cole SR, Kawachi I, Maller SR, Berkman LF. Test of item-response bias in the CES-D scale: Experience from the New Haven EPESE Study. Journal of Clinical Epidemiology. 2000; 53:285–289. DOI: 10.1016/S0895-4356(99)00151-1 [PubMed: 10760639]

Cook KF, Kallen MA, Amtmann D. Having a fit: Impact of number of items and distribution of data on traditional criteria for assessing IRT's unidimensionality assumption. Quality of Life Research. 2009; 18:447–460. DOI: 10.1007/s11136-009-9464-4 [PubMed: 19294529]

Cox DR, Snell EJ. The analysis of binary data. 2. London: Chapman and Hall; 1989.

Crane PK, Gibbons LE, Jolley L, van Belle G. Differential item functioning analysis with ordinal logistic regression techniques. DIFdetect and difwithpar. Medical Care. 2006; 44(11 Suppl 3):S115–S123. DOI: 10.1097/01.mlr.0000245183.28384.ed [PubMed: 17060818]

Crane PK, van Belle G, Larson EB. Test bias in a cognitive test: Differential item functioning in the CASI. Statistics in Medicine. 2004; 23:241–256. DOI: 10.1002/sim.1713 [PubMed: 14716726]

Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika. 1951; 16:297–334. DOI: 10.1007/BF02310555

Diener E, Suh EM, Lucas RE, Smith HL. Subjective well-being: Three decades of progress. Psychological Bulletin. 1999; 125(2):276–302. DOI: 10.1037/0033-2909.125.2.276

Diener E, Emmons RA, Larsen RJ, Griffin S. The satisfaction with life scale. Journal of Personality Assessment. 1985; 49(1):71–75. DOI: 10.1207/s15327752jpa4901_13 [PubMed: 16367493]

Dolan P, Peasgood T, White M. Do we really know what makes us happy? A review of the economic literature on the factors associated with subjective well-being. Journal of Economic Psychology. 2008; 29(1):94–122. DOI: 10.1016/j.joep.2007.09.001

Estabrook R, Sadler ME, McGue M. Differential item functioning in the Cambridge Mental Disorders in the Elderly (CAMDEX) Depression Scale across middle age and late life. Psychological Assessment. 2015; 27(4):1219–1233. DOI: 10.1037/pas0000114 [PubMed: 25938337]

Fleer PF. A Monte Carlo assessment of a new measure of item and test bias. Illinois Institute of Technology. Dissertation Abstracts International. 1993; 54(04B):2266.

Flowers CP, Oshima TC, Raju NS. A description and demonstration of the polytomous DFIT framework. Applied Psychological Measurement. 1999; 23:309–326. DOI: 10.1177/01466219922031437

Garrido LE, Abad FJ, Ponsoda V. A new look at Horn's parallel analysis with ordinal variables. Psychological Methods. 2012; 18:454–474. DOI: 10.1037/a0030005 [PubMed: 23046000]

Grayson DA, Mackinnon A, Jorm AF, Creasey H, Broe GA. Item bias in the Center for Epidemiologic Studies Depression Scale: Effects of physical disorders and disability in an elderly community sample. Journals of Gerontology: Psychological Sciences. 2000; 55B(5):273–282. DOI: 10.1093/geronb/55.5.P273

Green SB, Redell N, Thompson MS, Levy R. Accuracy of revised and traditional parallel analyses for assessing dimensionality with binary data. Educational and Psychologial Measurement. 2016; 76:5–21. DOI: 10.1177/0013164415581898

Gurland BJ, Cheng H, Maurer MS. Health-related restrictions of choices and choosing: Implications for quality of life and clinical interventions. Patient Related Outcome Measures. 2010; 1:73–80. DOI: 10.2147/PROM.S11842 [PubMed: 22915954]

Gurland BJ, Gurland RV. The choices, choosing model of quality of life: Description and rationale. International Journal of Geriatric Psychiatry. 2009a; 24(1):90–95. DOI: 10.1002/gps.2110 [PubMed: 18836984]

Gurland BJ, Gurland RV. The choices, choosing model of quality of life: Linkages to a science base. International Journal of Geriatric Psychiatry. 2009b; 24:84–89. DOI: 10.1002/gps.2109 [PubMed: 18836983]

Gurland BJ, Gurland R, Mitty E, Toner J. The choices, choosing model of quality of life: Clinical evaluation and intervention. Journal of Interprofessional Care, Informal Healthcare. 2009; 23(2): 110–120. DOI: 10.1080/13561820802675657

Gurland B, Teresi JA, Eimicke JP, Maurer MS, Reid MC. Quality of life impacts in the 16-year survival of an older ethnically diverse cohort. International Journal of Geriatric Psychiatry. 2014; 29:533–545. DOI: 10.1002/gps.4038 [PubMed: 24167085]

Hickey A, Barker M, McGee H, O'Boyle C. Measuring health-related quality of life in older patient populations: A review of current approaches. Pharmacoeconomics. 2005; 23(10):971–993. DOI: 10.2165/00019053-200523100-00002 [PubMed: 16235972]

Holmes D, Ory M, Teresi J. Special dementia care: Research, policy, and practice issues. Alzheimer's Disease and Associated Disorders: An International Journal. 1994; 8(Suppl 1)

Horn JL. A rationale and test for the number of factors in factor analysis. Psychometrika. 1965; 30:179–185. DOI: 10.1007/BF02289447 [PubMed: 14306381]

Iwata N, Turner RJ, Lloyd DA. Race/ethnicity and depressive symptoms in community-dwelling young adults: A differential item functioning analysis. Psychiatric Research. 2002; 110(3):281–289. DOI: 10.1016/S0165-1781(02)00102-6

Jensen RE, King-Kallimanis BL, Sexton E, Reeve BB, Moinpour CM, Potosky AL, et al. Measurement properties of PROMIS® sleep disturbance short forms in a large, ethnically diverse cancer cohort. Psychological Test and Assessment Modeling. 2016:58.

Kahneman D, Krueger AB. Developments in the measurement of subjective well-being. The Journal of Economic Perspectives. 2006; 20(1):23–24. DOI: 10.1257/089533006776526030

Kahneman D, Krueger AB, Schkade D, Shwarz N, Stone AA. Would you be happier if you were richer? A focusing illusion. Science. 2006; 312:1908–1910. DOI: 10.1126/science.1129688 [PubMed: 16809528]

Kapteyn A, Lee J, Tasscot C, Vonkova H, Zamarro G. Dimensions of subjective well-being. Social Indicators Research. 2015; 123(3):625–660. DOI: 10.1007/s11205-014-0753-0 [PubMed: 26316674]

Kim G, Chiriboga DA, Jang Y. Cultural equivalence in depressive symptoms in older White, Black, and Mexican-American adults. Journal of the American Geriatrics Society. 2009; 75(5):790–796. DOI: 10.1111/j.1532-5415.2009.02188.x

Kim S, Cohen AS, Alagoz C, Kim S. DIF detection and effect size measures for polytomously scored items. Journal of Educational Measurement. 2007; 44:93–116. DOI: 10.1111/j.1745-3984.2007.00029.x

Kim Y, Pilkonis PA, Frank E, Thase ME, Reynolds CF. Differential functioning of the Beck Depression Inventory in late-life patients: Use of item response theory. Psychology and Aging. 2002; 17(3):379–391. DOI: 10.1037/0882-7974.17.3.379 [PubMed: 12243380]

Kleinman M, Teresi JA. Differential item functioning magnitude and impact measures from item response theory models. Psychological Test and Assessment Modeling. 2016; 58:79–98. [PubMed: 28706769]

Kopf J, Zeileis A, Stobl C. Anchor selection strategies for DIF analysis: Review, assessment and new approaches. Educational and Psychological Measurement. 2015; 75:22–56. DOI: 10.1177/0013164414529792 [PubMed: 29795811]

Lawton MP. Assessing quality of life in Alzheimer disease research. Alzheimer Disease and Associated Disorders. 1997; 11(Suppl 6):91–99. [PubMed: 9437453]

Lawton MP. Environment and other determinants of well-being in older people. The Gerontologist. 1983; 23(4):349–357. DOI: 10.1093/geront/23.4.349 [PubMed: 6352420]

Lawton MP, Moss M, Hoffman C, Grant R, Ten Have T, Kleban MH. Health, valuation of life, and the wish to live. The Gerontologist. 1999; 39(4):406–416. DOI: 10.1093/geront/39.4.406 [PubMed: 10495578]

Lindwall M, Barkoukis V, Grano C, Lucidi G, Raudsapp L, Liukkonen J, et al. Method effects: The problem with negatively versus positively keyed items. Journal of Personality Assessment. 2012; 94(2):196–204. DOI: 10.1080/00223891.2011.645936 [PubMed: 22339312]

Lord FM. Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum; 1980.

Lord FM, Novick MR. Statistical theories of mental test scores. Reading Massachusetts: Addison-Wesley Publishing Company, Inc; 1968. (with contributions by A. Birnbaum)

Maydeu-Olivares A, Coffman DL. Random intercept item factor analysis. Psychological Methods. 2006; 11:344–362. DOI: 10.1037/1082-989X.11.4.344 [PubMed: 17154751]

McDonald RP. A basis for multidimensional item response theory. Applied Psychological Measurement. 2000; 24:99–114. DOI: 10.1177/01466210022031552

McDonald RP. Test theory: A unified treatment. Mahwah, NJ: L. Erlbaum Associates; 1999.

McFadden D. Conditional logit analysis of qualitative choice behavior. In: Zarembka P, editorFrontiers in Econometrics. New York: Academic Press; 1974. 105–142.

McHorney CA, Fleishman JA. Assessing and understanding measurement equivalence in health outcomes measures: Issues for further quantitative and qualitative inquiry. Medical Care. 2006; 44(Suppl 3):S205–S210. DOI: 10.1097/01.mlr.0000245451.67862.57 [PubMed: 17060829]

Meade AW, Johnson EC, Bradley PW. Power and sensitivity of alternative fit indices in tests of measurement invariance. Journal of Applied Psychology. 2008; 93:568–592. DOI: 10.1037/0021-9010.93.3.568 [PubMed: 18457487]

Meredith W. Measurement invariance, factor analysis and factorial invariance. Psychometrika. 1993; 58:525–543. DOI: 10.1007/BF02294825

Meredith W, Teresi JA. An essay on measurement and factorial invariance. Medical Care. 2006; 44(Suppl 3):S69–S77. DOI: 10.1097/01.mlr.0000245438.73837.89 [PubMed: 17060838]

Molloy DW, Standish TI. A guide to the standardized Mini-Mental Status Examination. International Psychogeriatrics. 1997; 9(Suppl 1):87–94. DOI: 10.1017/S1041610297004754 [PubMed: 9447431]

Mukherjee S, Gibbons LE, Kristiansson E, Crane PK. Extension of an iterative hybrid ordinal logistic regression/item response theory approach to detect and account for differential item functioning in longitudinal data. Psychological Test and Assessment Modeling. 2013; 55:127–147. [PubMed: 24432199]

Muthén LK, Muthén BO. M-PLUS users guide. 6. Los Angeles, California: Muthén and Muthén; 2011.

Nagelkerke NJD. A note on a general definition of the coefficient of determination. Biometrika. 1991; 78:691–692. DOI: 10.1093/biomet/78.3.691

Orlando-Edelen M, Thissen D, Teresi JA, Kleinman M, Ocepek-Welikson K. Identification of differential item functioning using item response theory and the likelihood-based model comparison approach: Applications to the Mini-Mental State Examination. Medical Care. 2006; 44(11 Suppl 3):S134–S142. DOI: 10.1097/01.mlr.0000245251.83359.8c [PubMed: 17060820]

Oshima TC, Kushubar S, Scott JC, Raju NS. DFIT for window user's manual: Differential functioning of items and tests. St. Paul, MN: Assessment Systems Corporation; 2009.

Perkins AJ, Stump TE, Monahan PO, McHorney CA. Assessment of differential item functioning for demographic comparisons in the MOS SF-36 health survey. Quality of Life Research. 2006; 15:331–348. [PubMed: 16547771]

Pickard AS, Dalal MR, Bushnell DM. A comparison of depressive symptoms in stroke and primary care: Applying Rasch models to evaluate the Center for Epidemiologic Studies-Depression Scale. Value in Health. 2006; 9(1):59–64. DOI: 10.1111/j.1524-4733.2006.00082.x [PubMed: 16441526]

Pilkonis PA, Choi SW, Reise SP, Stover AM, Riley WT, Cella D. Item banks for measuring emotional distress from the patient-reported outcomes measurement information system (PROMIS): Depression, anxiety, and anger. Assessment. 2011; 18:263–283. DOI: 10.1177/1073191111411667 [PubMed: 21697139]

R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: 2008.

Radloff LS. The CES-D scale: A self-report depression scale for research in the general population. Applied Psychological Measurement. 1977; 1:385–401. DOI: 10.1177/014662167700100306

Raju NS. DFITP5: A Fortran program for calculating dichotomous DIF/DTF [Computer program]. Chicago: Illinois Institute of Technology; 1999.

Raju NS, Fortmann-Johnson KA, Kim W, Morris SB, Nering ML, Oshima TC. The item parameter replication method for detecting differential functioning in the DFIT framework. Applied Measurement in Education. 2009; 33:133–147. DOI: 10.1177/0146621608319514

Raju NS, van der Linden WJ, Fleer PF. IRT-based internal measures of differential functioning of items and tests. Applied Psychological Measurement. 1995; 19:353–368. DOI: 10.1177/014662169501900405

Reeve BB, Hays RD, Bjorner JB, Cook KF, Crane PK, Teresi JA, et al. Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the Patient-Reported Outcome Measurement Information System (PROMIS). Medical Care. 2007; 45(5 Suppl 1):S22–S31. DOI: 10.1097/01.mlr.0000250483.85507.04 [PubMed: 17443115]

Reeve BB, Pinheiro LC, Jensen RE, Teresi JA, Potosky AL, McFatrich MK, et al. Psychometric evaluation of the PROMIS® fatigue measure in an ethnically and racially diverse population-based sample of cancer patients. Psychological Test and Assessment Modeling. 2016; 58(1):119–139.

Reise SP. The rediscovery of bifactor measurement models. Multivariate Behavioral Research. 2012; 47:667–696. DOI: 10.1080/00273171.2012.715555 [PubMed: 24049214]

Reise SP, Haviland MG. Item response theory and the measurement of clinical change. Journal of Personality Assessment. 2005; 84:228–238. DOI: 10.1207/s15327752jpa8403_02 [PubMed: 15907159]

Reise SP, Moore TM, Haviland MG. Bi-factor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. Journal of Personality Assessment. 2010; 92:544–559. DOI: 10.1080/00223891.2010.496477 [PubMed: 20954056]

Reise SP, Morizot J, Hays RD. The role of the bifactor model in resolving dimensionality issues in health outcomes measures. Quality of Life Research. 2007; 16(Suppl 1):19–31. DOI: 10.1007/s11136-007-9183-7 [PubMed: 17479357]

Reise SP, Widaman KF, Pugh RH. Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. Psychological Bulletin. 1993; 114:552–566. DOI: 10.1037/0033-2909.114.3.552 [PubMed: 8272470]

Revelle W. Psych: Package psych. 2015. http://CRAN.R-project.org/package=PSYCH

Revelle W, Zinbarg RE. Coefficient alpha, beta, omega, and the GLB: Comments on Sijtsma. Psychometrika. 2009; 74:145–154. DOI: 10.1007/s11336-008-9102-z

Rizopoulus D. Ltm: Latent Trait Models under IRT. 2009. https://cran.r-project.org/web/packages/ltm/index.html

Sass DA, Schmitt TA, Marsh HW. Evaluating model fit with ordered categorical data within a measurement invariance framework: A comparison of estimators. Structural Equation Modeling. 2014; 21:167–180.

Samejima F. Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph Supplement. 1969; 34:100–114. DOI: 10.1007/BF02290599

Saris WE, Revilla M, Krosnick JA, Shaeffer EM. Comparing questions with agree/disagree response options to questions with item-specific response options. Survey Research Methods. 2010; 4:61–79. DOI: 10.18148/srm/2010.v4i1.2682

Schmid L, Leiman J. The development of hierarchical factor solutions. Psychometrika. 1957; 22:53–61. DOI: 10.1007/BF02289209

Seligman MEP, Csikszentmihalyi M. Positive psychology: An introduction. American Psychologist. 2000; 55(1):5–14. DOI: 10.1037/0003-066X.55.1.5 [PubMed: 11392865]

Setodji CM, Reise SP, Morales LS, Fongwam N, Hays RD. Differential item functioning by survey language among older Hispanics enrolled in Medicare managed care a new method for anchor item selection. Medical Care. 2011; 49:461–468. DOI: 10.1097/MLR.0b013e318207edb5 [PubMed: 21422959]

Sijtsma K. On the use, the misuse, and the very limited usefulness of Cronbach's alpha. Psychometrika. 2009; 74:107–120. DOI: 10.1007/s11336-008-9101-0 [PubMed: 20037639]

Steptoe A, Deaton A, Stone AA. Subjective wellbeing, health and ageing. The Lancet. 2015; 385(9968):640–648. DOI: 10.1016/S0140-6736(13)61489-0

Steptoe A, Demakakos P, de Oliveira C, Wardle J. Distinctive biological correlates of positive psychological well-being in older men and women. Psychosomatic Medicine. 2012; 74:501–508. DOI: 10.1097/PSY.0b013e31824f82c8 [PubMed: 22511728]

Stone AA, Schwartz JE, Broderick JE, Deaton A. A snapshot of the age distribution of psychological well-being in the United States. Proceedings of the National Academy of Sciences of the United

States of America. 2010; 107:19949–19952. DOI: 10.1073/pnas.1003744107 [PubMed: 21041638]

Stout WF. A new item response theory modeling approach with applications to unidimensional assessment and ability estimation. Psychometrika. 1990; 55:293–325. DOI: 10.1007/BF02295289

Swaminathan H, Rogers HJ. Detecting differential item functioning using logistic regression procedures. Journal of Educational Measurement. 1990; 27:361–370. DOI: 10.1111/j.1745-3984.1990.tb00754.x

Teresi JA, Jones RN. Methodological issues in examining measurement equivalence in patient reported outcomes measures: Methods overview to the two-part series, “Measurement equivalence of the Patient Reported Outcomes Measurement Information System (PROMIS) short form measures”. Psychological Test and Assessment Modeling. 2016; 58:37–78. [PubMed: 28983448]

Teresi J, Abrams R, Holmes D. Measurement of depression and depression recognition individuals with cognitive impairment. In: Albert S, Logsdon R, editorsAssessing quality of life in Alzheimer’s disease. New York: Springer; 2000. 121–151.

Teresi JA, Kleinman M, Ocepek-Welikson K. Modern psychometric methods for detection of differential item functioning: Application to cognitive assessment measures. Statistics in Medicine. 2000; 19:1651–1683. DOI: 10.1002/(SICI)1097-0258(20000615/30)19:11/12<1651::AID-SIM453>3.0.CO;2-H [PubMed: 10844726]

Teresi J, Ocepek-Welikson K, Kleinman M, Eimicke JE, Crane PK, Jones RN, et al. Analysis of differential item functioning in the depression item bank from the Patient Reported Outcome Measurement Information System (PROMIS): An item response theory approach. Psychology Science Quarterly. 2009; 51:148–180. [PubMed: 20336180]

Teresi JA, Ocepek-Welikson K, Kleinman M, Ramirez M, Kim G. Psychometric properties and performance of the Patient Reported Outcomes Measurement Information System (PROMIS®) Depression short forms in ethnically diverse groups. Psychological Test and Assessment Modeling. 2016; 58:141–181. [PubMed: 28553573]

Teresi JA, Ramirez M, Lai J-S, Silver S. Occurrences and sources of Differential Item Functioning (DIF) in patient-reported outcome measures: Description of DIF methods, and review of measures of depression, quality of life and general health. Psychology Science Quarterly. 2008; 50:538–612. [PubMed: 20165561]

Thissen D, Steinberg L, Kuang D. Quick and easy implementation of the Benjamini-Hochberg procedure for controlling the false discovery rate in multiple comparisons. Journal of Educational and Behavioral Statistics. 2002; 27:77–83.

Thissen D, Steinberg L, Wainer H. Detection of differential item functioning using the parameters of item response models. In: Holland PW, Wainer H, editorsDifferential item functioning. Hillsdale, NJ: Lawrence Erlbaum, Inc; 1993. 123–135.

Toner JA, Teresi JA, Gurland B, Tirumalasetti F. The Feeling-Tone Questionnaire: Reliability and validity of a direct patient assessment screening instrument for detection of depressive symptoms in cases of dementia. Journal of Clinical Geropsychiatry. 1999; 5:63–78. DOI: 10.1023/A:1022994930394

Uebersax JS. Polycorr. A program for estimation of the standard and extended polychoric correlation coefficient. Computer program documentation manual. 2000

van Sonderen E, Sanderman R, Coyne JC. Ineffectiveness of reverse wording of questionnaire items: Let’s learn from cows in the rain. PloS ONE. 2013; 8(7)doi: 10.1371/journal.pone.0068967

Vecchione M, Allesandri G, Vittorio Caprara G, Tisak J. Are methods effects permanent or ephemeral in nature? The case of the revised life orientation test. Structural Equation Modeling. 2014; 21:117–130. DOI: 10.1080/10705511.2014.859511

Wainer H. Model-based standardization measurement of an item’s differential impact. In: Holland PW, Wainer H, editorsDifferential item functioning. Hillsdale, NJ: Lawrence Erlbaum, Inc; 1993. 123–135.

Wang WC, Shih CL, Sun GW. The DIF-free-then-DIF strategy for the assessment of differential item functioning. Educational and Psychological Measurement. 2012; 72:687–708. DOI: 10.1177/0013164411426157

Watson D, Clark LA, Tellegan A. Development and validation of brief measures of positive and negative affect: The PANAS scales. Journal of Personality and Social Psychology. 1988; 54(6): 1063–1070. DOI: 10.1037//0022-3514.54.6.1063 [PubMed: 3397865]

Wood AM, Taylor PJ, Joseph S. Does the CES-D measure a continuum from depression to happiness? Comparing substantive and artifactual models. Psychiatry Research. 2010; 177(1–2):120–123. DOI: 10.1016/j.psychres.2010.02.003 [PubMed: 20207424]

Woods CM. Empirical selection of anchors for tests of differential item functioning. Applied Psychological Measurement. 2009; 33:42–57. DOI: 10.1177/0146621607314044

Woods CM, Cai L, Wang M. The Langer-improved Wald test for DIF testing with multiple groups: Evaluation and comparison to two-group IRT. Educational and Psychological Measurement. 2013; 73:532–547. DOI: 10.1177/0013164412464875

Yang FM, Jones RN. Center of Epidemiologic Studies-Depression scale (CES-D) item response bias found with Mantel-Haenszel method was successfully replicated using latent variable modeling. Journal of Clinical Epidemiology. 2007; 60:1195–1200. DOI: 10.1016/j.jclinepi.2007.02.008 [PubMed: 17938063]

Yin L, Muramatsu N, Gordon R. Evaluating differential item functioning of a CES-D short form in Chinese older men and women: A rasch analysis. The Gerontologist. 2015; 55(Suppl 2):708.doi: 10.1093/geront/gnv355.12

Zinbarg RE, Revelle W, Yovel I, Li W. Cronbach's α, Revelle's β and McDonald's ω_H: Their relations with each other and two alternative conceptualizations of reliability. Psychometrika. 2005; 70(1): 123–133. http://personality-project.org/revelle/publications/zinbarg.revelle.pmet.05.pdf. DOI: 10.1007/s11336-003-0974-7

Zumbo BD. A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores. Ottawa, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense; 1999. http://www.educ.ubc.ca/faculty/zumbo/DIF/index.html

Zumbo BD, Gadermann AM, Zeisser C. Ordinal versions of coefficient alpha and theta for Likert rating scales. Journal of Modern Applied Statistical Methods. 2007; 6:21–29.

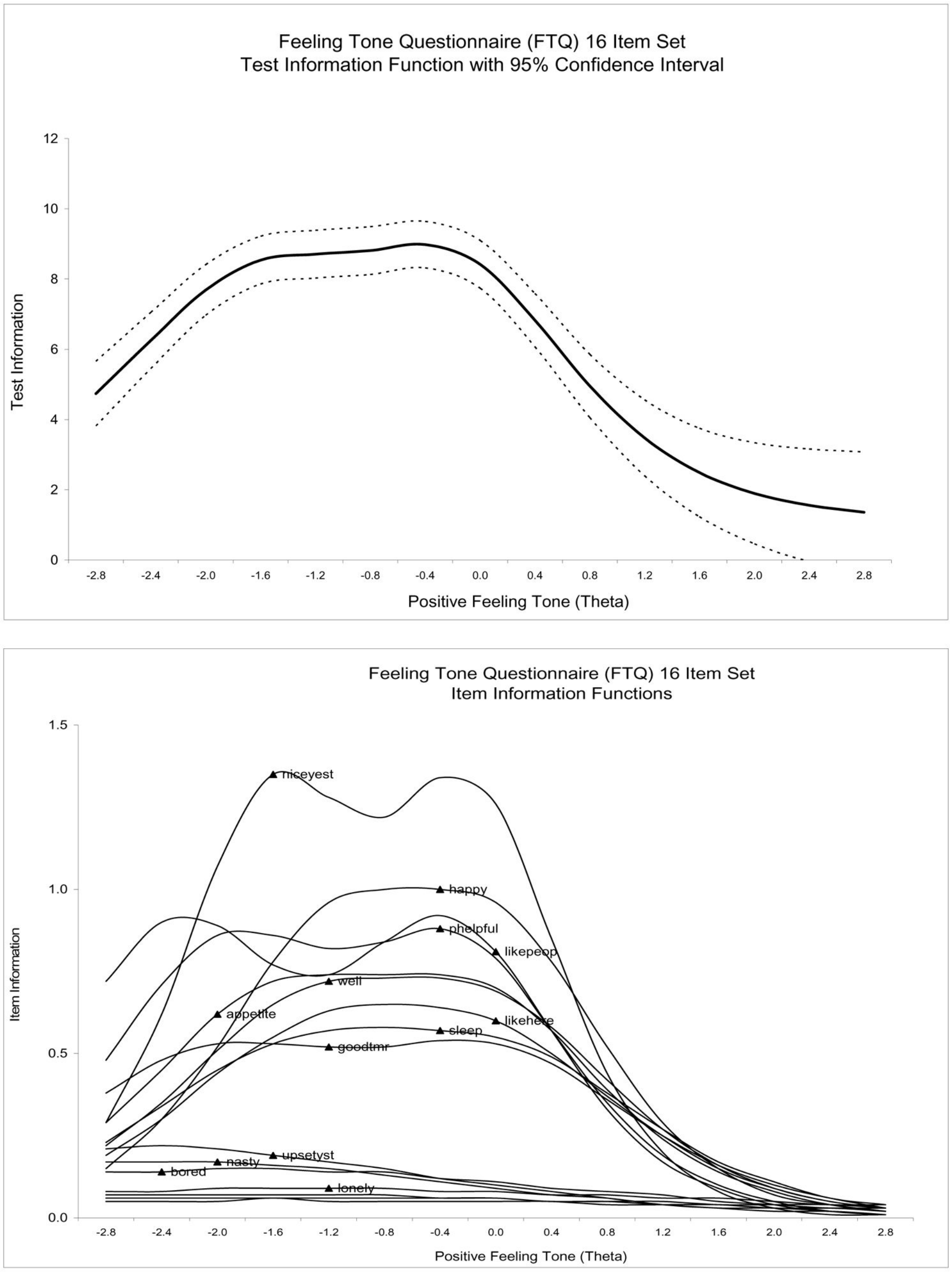


Figure 1.

Feeling Tone Questionnaire (FTQ) 16 item set: Test and item information functions

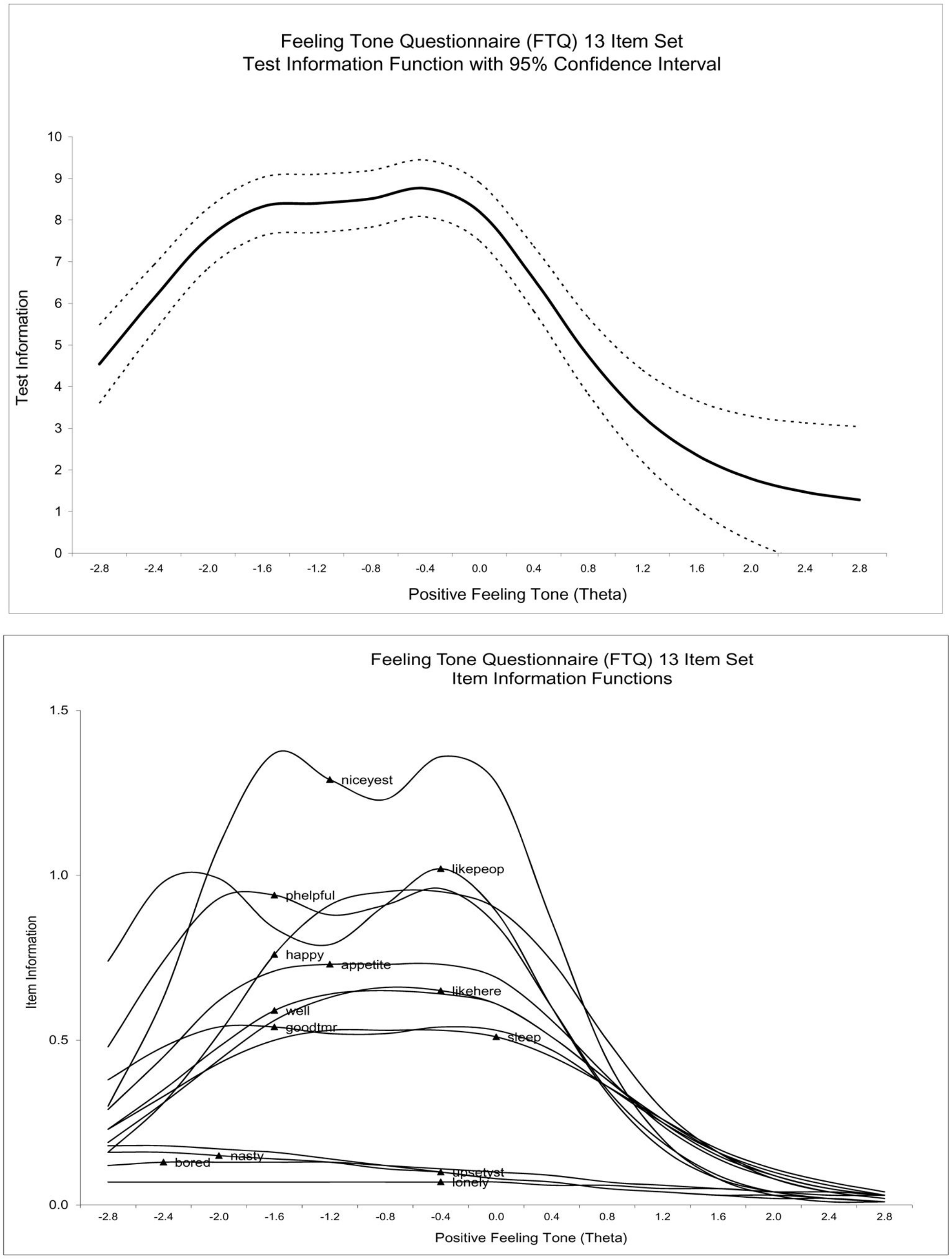


Figure 2.

Feeling Tone Questionnaire (FTQ) 13 item set: Test and item information functions

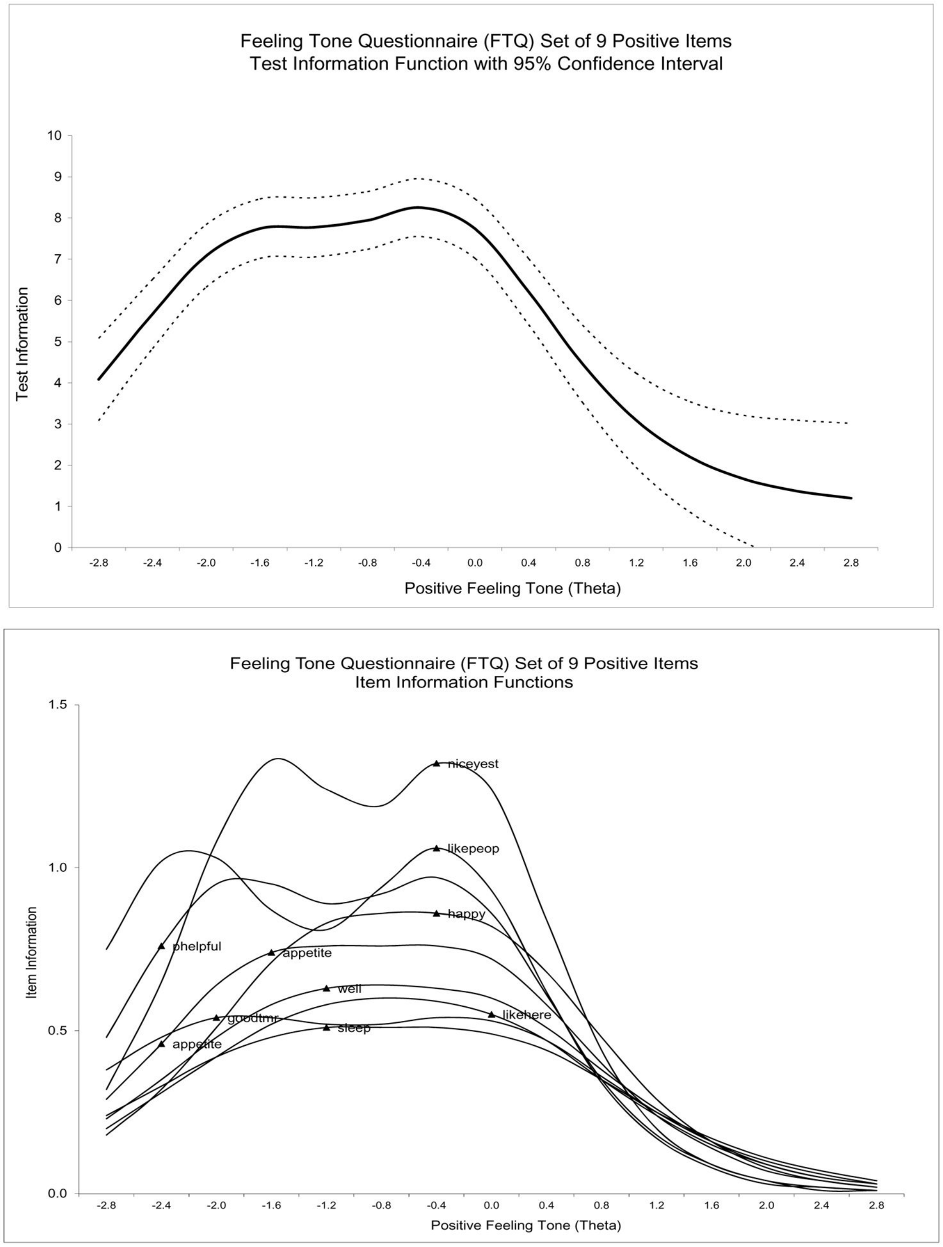


Figure 3.

Feeling Tone Questionnaire (FTQ) 9 positive affect item set: Test and item information functions

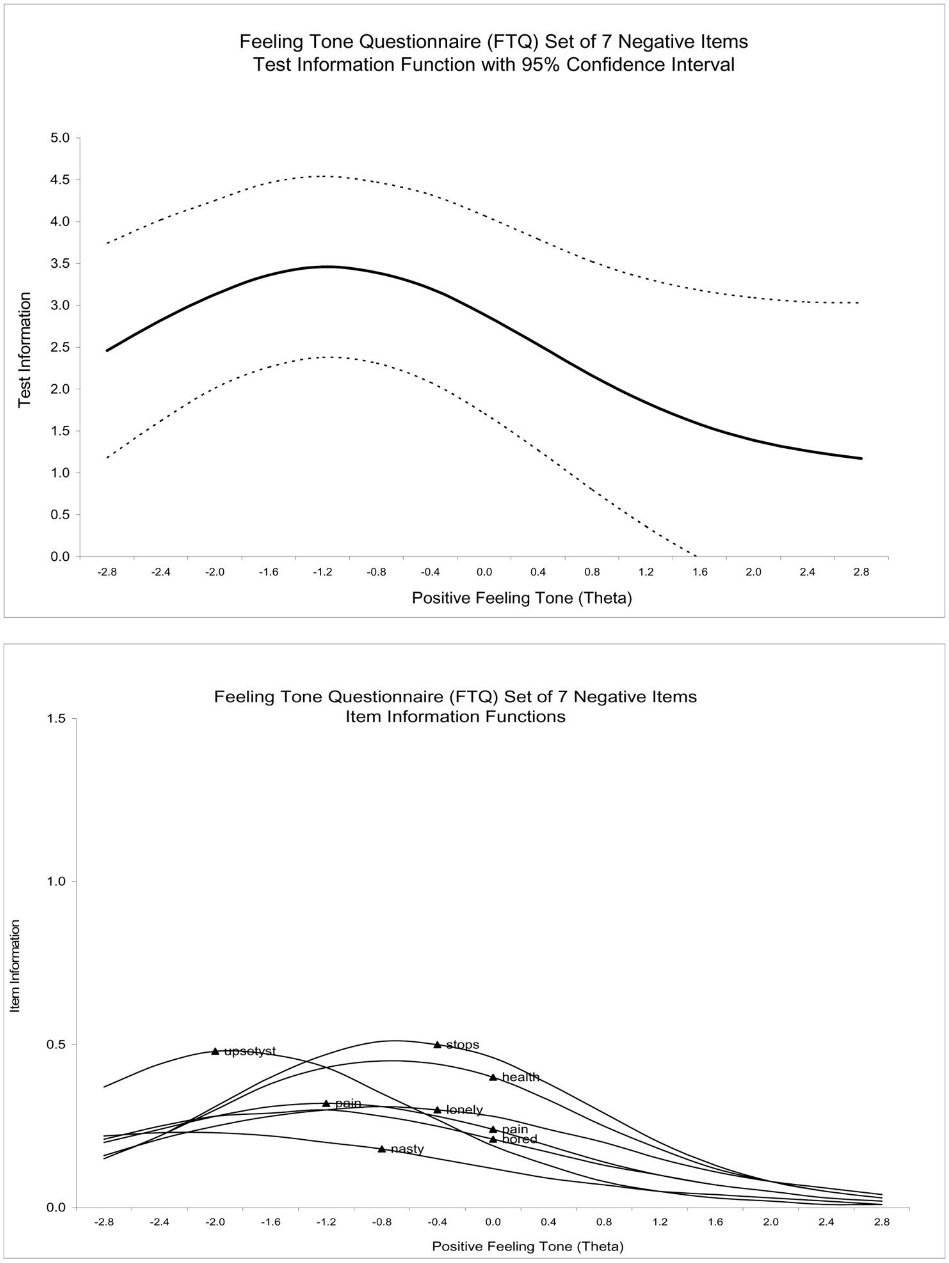
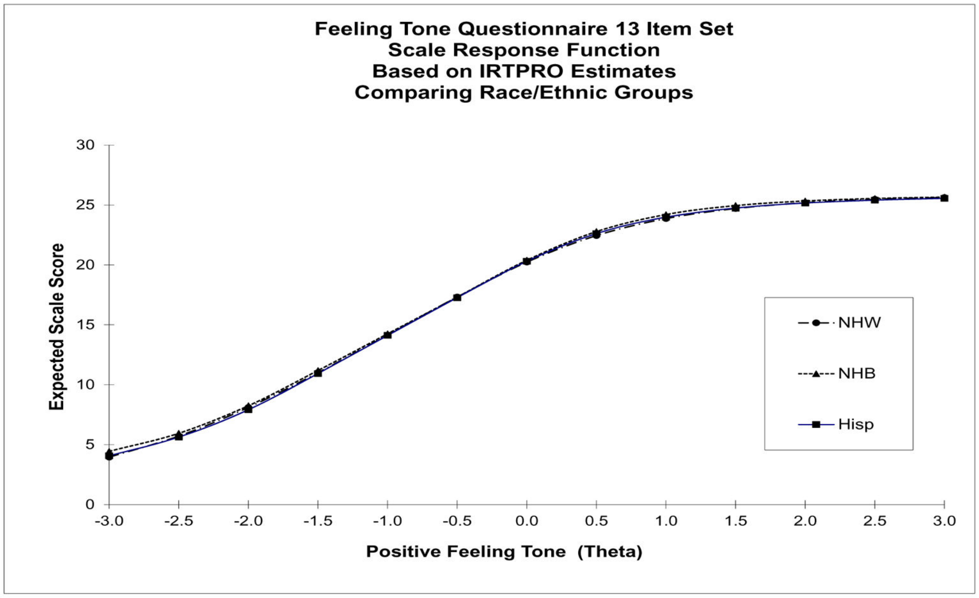
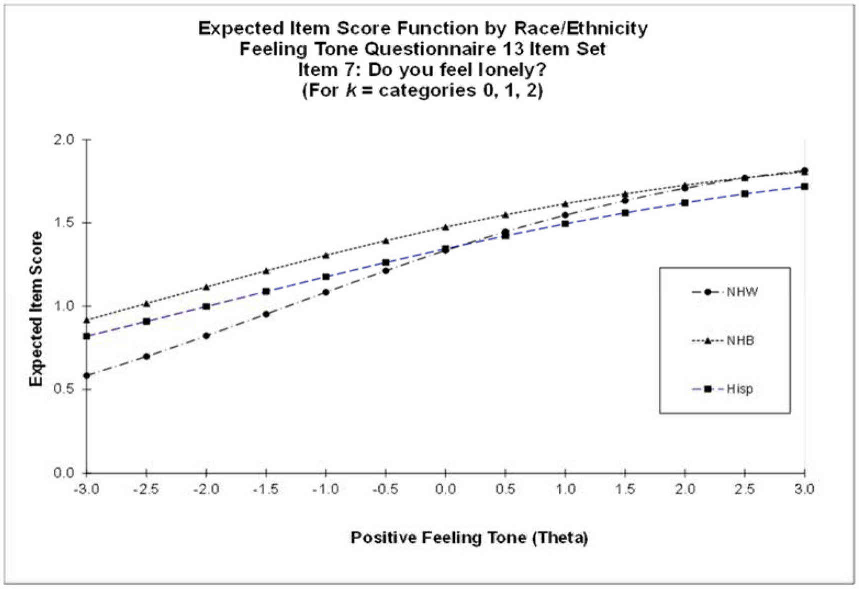


Figure 4.

Feeling Tone Questionnaire (FTQ) 7 negative affect item set: Test and item information functions

13 Item Set


Feeling Tone Questionnaire 13 Item Set
Scale Response Function
Based on IRTPRO Estimates
Comparing Race/Ethnic Groups
Expected Scale Score
30
25
20
15
10
5
0
-3.0
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Positive Feeling Tone (Theta)
NHW
NHB
Hisp


Expected Item Score Function by Race/Ethnicity
Feeling Tone Questionnaire 13 Item Set
Item 7: Do you feel lonely?
(For k = categories 0, 1, 2)
Expected Item Score
2.0
1.5
1.0
0.5
0.0
-3.0
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Positive Feeling Tone (Theta)
NHW
NHB
Hisp

9 Positive Affect Items Set

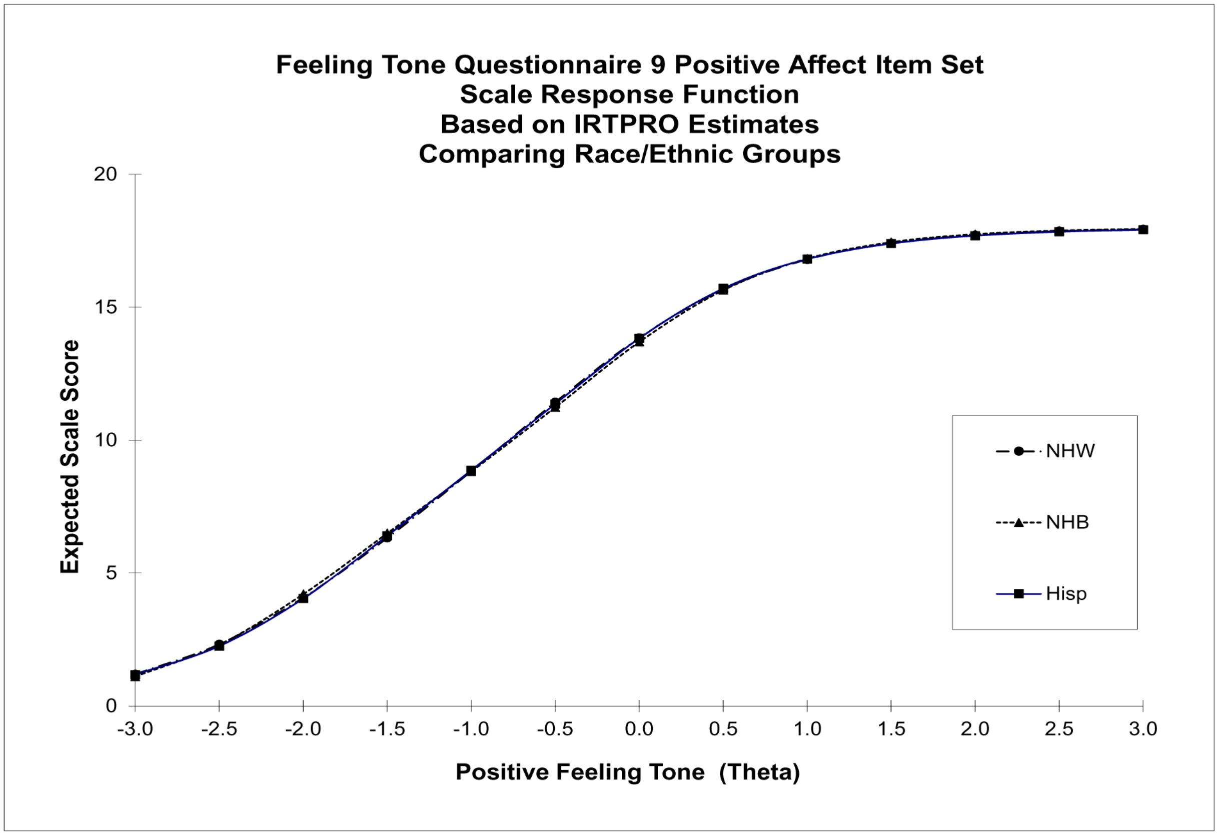


Figure 5.

Feeling Tone Questionnaire (FTQ): Expected scale and item scores. Race/ethnicity subgroups

Note: The curve for the non-Hispanic Black group is above that of the non-Hispanic White group across the level of theta, indicating a higher probability of responding to the loneliness item in the direction of not feeling lonely because the item was reverse-scored.

Table 1

Demographic characteristics of the sample by race/ethnicity and total sample

	Non-Hispanic White	Non-Hispanic Black	Hispanic	Total
N	4,960	1,144	517	6,621
Gender				
Male	1,306 (26%)	311 (27%)	176 (34%)	1,793 (27%)
Female	3,654 (74%)	832 (73%)	341 (66%)	4,827 (73%)
Age				
< 65	221 (5%)	174 (15%)	74 (14%)	469 (7%)
65 to 74	479 (10%)	254 (22%)	113 (22%)	846 (13%)
75 to 84	1,499 (30%)	354 (31%)	175 (34%)	2,028 (31%)
85 to 94	2,287 (46%)	304 (27%)	137 (27%)	2,728 (41%)
> 94	468 (9%)	54 (5%)	17 (3%)	539 (8%)
Average	83.9 (9.9)	77.3 (12.6)	82.2 (12.2)	82.2 (11.0)
Education				
0 years	597 (13%)	139 (14%)	79 (18%)	815 (14%)
1 to 8 years	909 (20%)	322 (32%)	202 (45%)	1,433 (24%)
9 to 12 years	1,996 (44%)	426 (42%)	126 (28%)	2,548 (42%)
> 12 years	1,051 (23%)	134 (13%)	38 (9%)	1,223 (20%)
0 to 11 years	1,975 (43%)	606 (59%)	338 (76%)	2,919 (49%)
12 & above	2,578 (57%)	415 (41%)	107 (24%)	3,100 (51%)
Average (s.d.)	9.9 (5.2)	8.6 (4.9)	6.6 (4.9)	9.4 (5.2)
Sample Type				
Nursing home	4,267 (86%)	857 (75%)	386 (75%)	5,510 (83%)
Assisted living	338 (7%)	36 (3%)	23 (4%)	397 (6%)
Community	355 (7%)	251 (22%)	108 (21%)	714 (11%)

Note: The sample sizes for some of the analyses presented in other tables are greater than that used in the analyses of race/ethnicity because there are some respondents who are in an “other” category for race/ethnicity. Due to missing data not all variable level sample sizes sum to the total sample size.

Table 2

Tests of dimensionality for the Feeling Tone Questionnaire from principal components analysis: Eigenvalues by subgroup

					Ratio
Statistic	**Component 1**	**Component 2**	**Component 3**	**Component 4**	**Component 1/Component 2**
Total Sample, 16 Item Set (n = 6756)					
Eigenvalues	5.200	2.232	1.291	0.951	2.3
Explained Variance	32.5%	14.0%	8.1%	5.9%	
First Random Half, 16 Item Set (n = 3378)					
Eigenvalues	5.186	2.200	1.324	0.908	2.4
Explained Variance	32.4%	13.8%	8.3%	5.7%	
Total Sample, 13 Item Set (n = 6756)					
Eigenvalues	4.995	1.577	0.950	0.890	3.2
Explained Variance	38.4%	12.1%	7.3%	6.8%	
First Random Half Sample, 13 Item Set (n = 3378)					
Eigenvalues	4.974	1.575	0.930	0.871	3.2
Explained Variance	38.3%	12.1%	7.2%	6.7%	
Total Sample, 9 Positive Affect Item Set (n = 6756)					
Eigenvalues	4.573	0.856	0.701	0.607	5.3
Explained Variance	50.8%	9.5%	7.8%	6.7%	
Total Sample, 7 Negative Affect Item Set (n = 6756)					
Eigenvalues	2.702	1.109	0.904	0.686	2.4
Explained Variance	38.6%	15.8%	12.9%	9.8%	
Non-Hispanic White - 16 Item Set (n = 4960)					
Eigenvalues	5.312	2.082	1.286	0.945	2.6
Explained Variance	33.2%	13.0%	8.0%	5.9%	
Non-Hispanic Black - 16 Item Set (n = 1144)					
Eigenvalues	4.863	2.551	1.367	1.039	1.9
Explained Variance	30.4%	15.9%	8.5%	6.5%	
Hispanic - 16 Item Set (n = 517)					
Eigenvalues	4.832	2.953	1.311	0.963	1.6

					Ratio
Statistic	**Component 1**	**Component 2**	**Component 3**	**Component 4**	**Component 1/Component 2**
Explained Variance	30.2%	18.5%	8.2%	6.0%	
Non-Hispanic White - 13 Item Set (n = 4960)					
Eigenvalues	5.047	1.512	0.950	0.893	3.3
Explained Variance	38.8%	11.6%	7.3%	6.9%	
Non-Hispanic Black - 13 Item Set (n = 1144)					
Eigenvalues	4.827	1.675	1.047	0.939	2.9
Explained Variance	37.1%	12.9%	8.1%	7.2%	
Hispanic - 13 Item Set (n = 517)					
Eigenvalues	4.728	1.829	1.077	0.943	2.6
Explained Variance	36.4%	14.1%	8.3%	7.3%	
Non-Hispanic White – 9 Positive Affect Item Set (n = 4960)					
Eigenvalues	4.563	0.849	0.722	0.593	5.4
Explained Variance	50.7%	9.4%	8.0%	6.6%	
Non-Hispanic Black - 9 Positive Affect Item Set (n = 1144)					
Eigenvalues	4.540	0.931	0.752	0.615	4.9
Explained Variance	50.4%	10.3%	8.4%	6.8%	
Hispanic - 9 Positive Affect Item Set (n = 517)					
Eigenvalues	4.361	1.017	0.745	0.697	4.3
Explained Variance	48.5%	11.3%	8.3%	7.7%	
Non-Hispanic White - 7 Negative Affect Item Set (n = 4960)					
Eigenvalues	2.692	1.128	0.872	0.694	2.4
Explained Variance	38.5%	16.1%	12.5%	9.9%	
Non-Hispanic Black - 7 Negative Affect Item Set (n = 1144)					
Eigenvalues	2.620	1.207	0.900	0.685	2.2
Explained Variance	37.4%	17.2%	12.9%	9.8%	
Hispanic - 7 Negative Affect Item Set (n = 517)					
Eigenvalues	3.095	1.158	0.859	0.648	2.7
Explained Variance	44.2%	16.5%	12.3%	9.3%	

Table 3

Item factor loadings (λ) for the Feeling Tone Questionnaire from the Mplus confirmatory factor analysis methods models comparison (Total sample n=6756)

Item Description	Unidimensional CFA Models				Models with Correlated Item Uniqueness		Negative Items Methods Factor Model (3)		BiFactor Two Group Factor Model Modified Model (4)			Two Factor Model (5)	
	16 Item set (1a)	13 Item Set (1b)	Positive Affect Items (1c)	Negative Affect Items (1d)	Correlated Negative (2a)	Correlated Positive & Negative (2b)							
	λ (s.e.)	λ (s.e.)	λ (s.e.)	λ (s.e.)	λ (s.e.)	λ (s.e.)	G λ (s.e.)	FM λ (s.e.)	G λ (s.e.)	F1 λ (s.e.)	F2 λ (s.e.)	F1 λ (s.e.)	F2 λ (s.e.)
Are you feeling well?	0.65 (0.01)	0.64 (0.01)	0.64 (0.01)		0.66 (0.01)	0.78 (0.03)	0.65 (0.01)		0.44 (0.02)	0.49 (0.01)		0.66 (0.01)	
Do you like being here?	0.63 (0.01)	0.64 (0.01)	0.63 (0.01)		0.64 (0.01)	0.68 (0.03)	0.64 (0.01)		0.34 (0.02)	0.53 (0.01)		0.63 (0.01)	
Are you feeling happy today?	0.71 (0.01)	0.71 (0.01)	0.70 (0.01)		0.72 (0.01)	0.84 (0.02)	0.72 (0.01)		0.48 (0.02)	0.54 (0.01)		0.72 (0.01)	
Do you have any pain?	0.29 (0.01)			0.53 (0.02)	0.21 (0.01)	0.25 (0.02)	0.20 (0.01)	0.51 (0.02)	0.40 (0.02)		0.41 (0.03)		0.51 (0.02)
Are you feeling bored?	0.37 (0.01)	0.35 (0.01)		0.51 (0.02)	0.30 (0.01)	0.32 (0.02)	0.31 (0.01)	0.42 (0.02)	0.61 (0.02)		−0.06 (0.03)		0.59 (0.02)
Are people helpful here?	0.69 (0.01)	0.71 (0.01)	0.72 (0.01)		0.70 (0.01)	0.42 (0.03)	0.70 (0.01)		0.16 (0.02)	0.73 (0.01)		0.70 (0.01)	
Do you feel lonely?	0.32 (0.01)	0.29 (0.01)		0.53 (0.02)	0.23 (0.01)	0.25 (0.02)	0.24 (0.01)	0.46 (0.02)	0.54 (0.02)		0.03 (0.02)		0.54 (0.02)
Is there anything that stops you from doing what you want to do?	0.26 (0.01)			0.60 (0.01)	0.15 (0.01)	0.19 (0.02)	0.15 (0.01)	0.59 (0.02)	0.43 (0.02)		0.32 (0.03)		0.51 (0.02)
Do you have a good appetite?	0.65 (0.01)	0.66 (0.01)	0.67 (0.01)		0.66 (0.01)	0.45 (0.03)	0.66 (0.01)		0.29 (0.02)	0.59 (0.01)		0.66 (0.01)	
Do you like the people here?	0.71 (0.01)	0.72 (0.01)	0.73 (0.01)		0.72 (0.01)	0.36 (0.03)	0.72 (0.01)		0.14 (0.02)	0.77 (0.01)		0.72 (0.01)	
Is anybody nasty to you?	0.36 (0.02)	0.34 (0.02)		0.44 (0.02)	0.30 (0.02)	0.32 (0.02)	0.31 (0.02)	0.34 (0.02)	0.52 (0.02)		0.03 (0.03)		0.54 (0.02)
Do you sleep well?	0.61 (0.01)	0.60 (0.01)	0.60 (0.01)		0.61 (0.01)	0.71 (0.03)	0.61 (0.01)		0.39 (0.02)	0.47 (0.01)		0.61 (0.01)	
Did you have a nice day yesterday?	0.74 (0.01)	0.75 (0.01)	0.75 (0.01)		0.75 (0.01)	0.66 (0.02)	0.75 (0.01)		0.36 (0.02)	0.66 (0.01)		0.75 (0.01)	
Did anything upset you yesterday?	0.39 (0.01)	0.35 (0.01)		0.56 (0.02)	0.31 (0.01)	0.34 (0.02)	0.32 (0.01)	0.46 (0.02)	0.59 (0.02)		0.12 (0.03)		0.62 (0.02)
Do you have any trouble with your health?	0.26 (0.01)			0.61 (0.01)	0.15 (0.01)	0.19 (0.02)	0.15 (0.01)	0.62 (0.02)	0.36 (0.03)		0.82 (0.05)		0.52 (0.02)
Do you feel good about tomorrow?	0.60 (0.01)	0.61 (0.01)	0.61 (0.01)		0.61 (0.01)	0.52 (0.03)	0.61 (0.01)		0.30 (0.02)	0.53 (0.01)		0.61 (0.01)	
CFI	0.829	0.925	0.967	0.876	0.940	0.973	0.926		0.950			0.920	
RMSEA	0.105 (0.103 – 0.107)	0.084 (0.082 – 0.087)	0.081 (0.077 – 0.085)	0.088 (0.083 – 0.094)	0.070 (0.067 – 0.072)	0.061 (0.058 – 0.064)	0.072 (0.070 – 0.074)		0.062 (0.060 – 0.064)			0.072 (0.070 – 0.074)	

Note: The uniqueness correlation between the items, nice day yesterday and feel good about tomorrow in Model 2b was set to zero

Table 4

Internal consistency and dimensionality estimates for the Feeling Tone Questionnaire: Alpha, McDonald’s Omega Total and explained common variance (ECV) for the total sample and demographic subgroups from the bifactor model analyses (from “Psych” R package)

Sample	N	Cronbach’s Alpha	Ordinal Alpha	McDonald’s Omega Total	ECV (2 group model)	ECV (3 group model)
16 item set						
Total Sample	6756	0.773	0.852	0.856	27.555	26.337
Random sample group 1	3378	0.773	0.852	0.856	28.331	27.114
Random sample group 2	3378	0.774	0.851	0.856	26.518	27.723
Race/Ethnicity: White	4960	0.779	0.858	0.862	33.944	30.734
Race/Ethnicity: Black	1144	0.743	0.827	0.835	7.201	23.790
Race/Ethnicity: Hispanic	517	0.757	0.834	0.839	12.025	33.516
13 Item Set						
Total Sample	6756	0.782	0.857	0.862	40.882	68.706
Random sample group 1	3378	0.781	0.858	0.862	42.231	67.890
Random sample group 2	3378	0.782	0.857	0.862	38.633	68.346
Race/Ethnicity: White	4960	0.784	0.861	0.865	43.935	70.614
Race/Ethnicity: Black	1144	0.771	0.845	0.852	25.874	34.049
Race/Ethnicity: Hispanic	517	0.769	0.842	0.848	26.481	59.743
Positive Affect Item Set						
Total Sample	6756	0.808	0.878	0.879	77.350	72.565
Random sample group 1	3378	0.803	0.875	0.875	76.619	72.859
Random sample group 2	3378	0.814	0.882	0.882	76.049	71.979
Race/Ethnicity: White	4960	0.804	0.878	0.878	77.847	73.145
Race/Ethnicity: Black	1144	0.815	0.876	0.877	71.934	65.772
Race/Ethnicity: Hispanic	517	0.803	0.864	0.866	68.854	67.783
Negative Affect Item Set						
Total Sample	6756	0.605	0.733	0.734	47.195	51.626
Random sample group 1	3378	0.608	0.737	0.737	44.079	47.600
Random sample group 2	3378	0.602	0.729	0.731	51.850	47.312
Race/Ethnicity: White	4960	0.601	0.732	0.733	44.680	53.227

Sample	N	Cronbach’s Alpha	Ordinal Alpha	McDonald’s Omega Total	ECV (2 group model)	ECV (3 group model)
Race/Ethnicity: Black	1144	0.591	0.719	0.721	47.186	34.854
Race/Ethnicity: Hispanic	517	0.677	0.782	0.787	56.692	48.505

Table 5

Item response theory (IRT) reliability estimates for the Feeling Tone Questionnaire (FTQ) 13 item sets at varying levels of the attribute (theta) estimate based on results of the IRT analysis (IRTPRO) for total sample and race/ethnic subgroups

FTQ (Theta)	13 item set			
	Total	Non-Hispanic White	Non-Hispanic Black	Hispanic
−2.8	0.82	0.83	0.81	0.79
−2.4	0.86	0.86	0.86	0.84
−2.0	0.89	0.88	0.89	0.87
−1.6	0.89	0.89	0.89	0.90
−1.2	0.89	0.90	0.89	0.90
−0.8	0.90	0.90	0.89	0.90
−0.4	0.90	0.90	0.89	0.90
0.0	0.89	0.89	0.90	0.90
0.4	0.87	0.86	0.89	0.89
0.8	0.83	0.81	0.85	0.86
1.2	0.77	0.75	0.80	0.81
1.6	0.70	0.69	0.74	0.74
Overall (Average)	0.85	0.85	0.86	0.86

Note: Reliability estimates are calculated for theta levels for which there are respondents

Table 6

Item response theory (IRT) reliability estimates for the Feeling Tone Questionnaire (FTQ) positive and negative affect item sets at varying levels of the attribute (theta) estimate based on results of the IRT analysis (IRTPRO) for total sample and race/ethnic subgroups

FTQ (Theta)	Positive affect items set				Negative affect items set			
	Total	**Non-Hispanic White**	**Non-Hispanic Black**	**Hispanic**	**Total**	**Non-Hispanic White**	**Non-Hispanic Black**	**Hispanic**
−2.8	0.80	0.81	0.80	0.76	0.71	N/A	0.69	N/A
−2.4	0.85	0.85	0.86	0.83	0.74	0.74	0.72	0.75
−2.0	0.87	0.87	0.88	0.87	0.76	0.76	0.75	0.79
−1.6	0.89	0.89	0.89	0.89	0.77	0.77	0.77	0.81
−1.2	0.89	0.89	0.89	0.89	0.77	0.77	0.77	0.83
−0.8	0.89	0.89	0.88	0.89	0.77	0.77	0.78	0.83
−0.4	0.89	0.89	0.89	0.89	0.76	0.75	0.77	0.82
0.0	0.89	0.88	0.90	0.90	0.74	0.74	0.75	0.81
0.4	0.86	0.85	0.88	0.89	0.72	0.71	0.72	0.77
0.8	0.82	0.81	0.85	0.86	0.68	0.68	0.68	0.73
1.2	0.75	0.74	0.79	0.80	0.65	0.65	0.64	0.68
1.6	0.69	0.67	0.73	0.74	N/A	N/A	N/A	N/A
Overall (Average)	0.84	0.84	0.85	0.85	0.73	0.73	0.73	0.78

Note: Reliability estimates are calculated for theta levels for which there are respondents

Table 7

Item response theory (IRT) item parameters and standard error estimates (using IRTPRO) for the Feeling Tone Questionnaire 13 and 9 positive affect item sets for the total sample (n = 6756)

Item Description	13 item set			Positive items set		
	a **(s.e. of *a*)**	***b1*** **(s.e.)**	***b2*** **(s.e.)**	***a*** **(s.e. of *a*)**	***b1*** **(s.e.)**	***b2*** **(s.e.)**
Are you feeling well?	1.50 (0.04)	−1.44 (0.04)	−0.06 (0.02)	1.49 (0.04)	−1.45 (0.04)	−0.06 (0.02)
Do you like being here?	1.50 (0.04)	−1.30 (0.03)	−0.09 (0.02)	1.43 (0.04)	−1.34 (0.04)	−0.09 (0.02)
Are you feeling happy today?	1.84 (0.05)	−1.21 (0.03)	−0.02 (0.02)	1.75 (0.05)	−1.25 (0.03)	−0.02 (0.02)
Do you have any pain?	Item excluded					
Are you feeling bored?	0.68 (0.03)	−2.53 (0.12)	−1.29 (0.07)			
Are people helpful here?	1.91 (0.06)	−1.91 (0.04)	−0.30 (0.02)	1.93 (0.06)	−1.91 (0.04)	−0.30 (0.02)
Do you feel lonely?	0.50 (0.03)	−2.49 (0.14)	−0.56 (0.06)			
Is there anything that stops you from doing what you want to do?	Item excluded					
Do you have a good appetite?	1.63 (0.05)	−1.61 (0.04)	−0.14 (0.02)	1.67 (0.05)	−1.60 (0.04)	−0.14 (0.02)
Do you like the people here?	1.99 (0.06)	−2.24 (0.05)	−0.34 (0.02)	2.04 (0.06)	−2.23 (0.05)	−0.34 (0.02)
Is anybody nasty to you?	0.74 (0.04)	−3.14 (0.16)	−2.02 (0.10)			
Do you sleep well?	1.37 (0.04)	−1.51 (0.04)	0.02 (0.02)	1.34 (0.04)	−1.53 (0.04)	0.02 (0.02)
Did you have a nice day yesterday?	2.30 (0.07)	−1.60 (0.03)	−0.20 (0.02)	2.26 (0.06)	−1.63 (0.03)	−0.20 (0.02)
Did anything upset you yesterday?	0.79 (0.04)	−3.20 (0.16)	−2.10 (0.10)			
Do you have any trouble with your health?	Item excluded					
Do you feel good about tomorrow?	1.43 (0.04)	−2.01 (0.05)	−0.04 (0.02)	1.43 (0.04)	−2.02 (0.05)	−0.04 (0.02)

Table 8

Feeling Tone Questionnaire (FTQ) 13 item set: Differential item functioning (DIF) results Race/Ethnicity subgroup comparisons

Item Description	IRTPRO		*lordif*		Magnitude (NCDIF)		Effect Size T1	
	NHW vs. NHB	NHW vs. Hispanic	NHW vs. NHB	NHW vs. Hispanic	NHW vs. NHB	NHW vs. Hispanic	NHW vs. NHB	NHW vs. Hispanic
Are you feeling well?	U^*	U^*	NU; U	NU^*; U	0.0041	0.0143	0.0613	**0.1066**†
Do you like being here?					0.0013	0.0002	0.0121	−0.0020
Are you feeling happy today?	U	U^*	U	NU; U^*	0.0028	0.0120	−0.0518	**−0.1025**†
Do you have any pain?	Item excluded							
Are you feeling bored?	U		U^*		0.0133	0.0015	**−0.1068**†	−0.0301
Are people helpful here?	U^*	U	U^*	U	0.0089	0.0063	0.0822	0.0712
Do you feel lonely?	NU; U^*		NU; U^*	NU	**0.0297**	0.0050	**−0.1622**†	−0.0426
Is there anything that stops you from doing what you want to do?	Item excluded							
Do you have a good appetite?	U^*	U^*	NU; U^*		0.0111	0.0064	0.0984	0.0579
Do you like the people here?	U^*		NU; U^*	NU	0.0088	0.0060	0.0706	0.0582
Is anybody nasty to you?			U		0.0021	0.0031	−0.0455	−0.0494
Do you sleep well?	U^*	U^*			0.0009	0.0103	−0.0090	0.0044
Did you have a nice day yesterday?	U^*	U	U		0.0035	0.0013	−0.0444	0.0125
Did anything upset you yesterday?	U^*	U	NU^*; U^*	U^*	0.0137	0.0132	−0.0970	**−0.1071**†
Do you have any trouble with your health?	Item excluded							
Do you feel good about tomorrow?					0.0015	0.0006	−0.0279	0.0231

Item 7 – ‘Lonely’ had NCDIF values larger than the threshold (0.0240)

[*] Asterisks indicate significance after adjustment for multiple comparisons.

[†] Indicates value above threshold of 0.10.

NU= Non-uniform DIF involving the discrimination parameters; U=Uniform DIF involving the location parameters.

For the *lordif* analyses, uniform and non-uniform DIF was determined using the likelihood ratio chi-square test. Uniform DIF is obtained by comparing the log likelihood values from models one and two. Non-uniform DIF is obtained by comparing the log likelihood values from models two and three. DIF was not detected using the pseudo R^2 measures of Cox & Snell (1989), Nagelkerke (1991), and McFadden (1974) or with the change in Beta criterion.

Table 9

Feeling Tone Questionnaire (FTQ) positive affect item set: Differential item functioning (DIF) results. Race/Ethnicity subgroup comparisons

Item Description	IRTPRO		*lordif*		Magnitude (NCDIF)		Effect Size T1	
	NHW vs. NHB	NHW vs. Hispanic	NHW vs. NHB	NHW vs. Hispanic	NHW vs. NHB	NHW vs. Hispanic	NHW vs. NHB	NHW vs. Hispanic
Positive affect item set								
Are you feeling well?	U*	U*	NU	NU*; U	0.0020	0.0116	0.0360	0.0794
Do you like being here?					0.0001	0.0015	−0.0107	−0.0254
Are you feeling happy today?	U*	NU; U*	U*	NU; U*	0.0076	0.0171	−0.0777	**−0.1264**†
Are people helpful here?	U*			NU	0.0036	0.0045	0.0540	0.0422
Do you have a good appetite?	U*	NU; U*	NU; U*		0.0059	0.0019	0.0706	0.0287
Do you like the people here?	U*			NU*	0.0037	0.0024	0.0460	0.0321
Do you sleep well?	U*	NU*; U*	U		0.0009	0.0067	−0.0293	−0.0174
Did you have a nice day yesterday?	U*	NU; U			0.0007	0.0006	0.0136	−0.0167
Do you feel good about tomorrow?	U*		U*		0.0022	0.0006	−0.0464	0.0022

*Asterisks indicate significance after adjustment for multiple comparisons.

†Indicates value above threshold of 0.10.

NU= Non-uniform DIF involving the discrimination parameters; U=Uniform DIF involving the location parameters.

For the *lordif* analyses, uniform and non-uniform DIF was determined using the likelihood ratio chi-square test. Uniform DIF is obtained by comparing the log likelihood values from models one and two. Non-uniform DIF is obtained by comparing the log likelihood values from models two and three. DIF was not detected using the pseudo R^2 measures of Cox & Snell (1989), Nagelkerke (1991), and McFadden (1974) or with the change in Beta criterion.